An Introduction to
MODERN STATISTICAL METHODS

BY

PAUL R. RIDER
Washington University
Saint Louis

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED
1939

PRINTED IN U. S. A.

PRESS OF
BRAUNWORTH & CO., INC
BUILDERS OF BOOKS
BRIDGEPORT CONN

PREFACE

In the past three decades enormous advances have been made in the field of statistics. These advances have taken place so rapidly that it is not at all surprising that those employing statistical methods have found it difficult to keep pace with them, or that certain of the older methods, which are obsolete, and even in some cases erroneous—or, at the very best, crudely approximative in character—continue to be taught in the classroom and to be treated in textbooks which appear from time to time. A notable example is the use of the probable error or standard error in testing the significance of a correlation coefficient derived from a sample, although this method gives unreliable or incorrect results if there is a high degree of correlation in the population from which the sample is drawn, or if the number in the sample is small.

It is, of course, in the theory of small samples that the greatest progress has been made. The theory of sampling which assumes that the sample is composed of a large number of items is inadequate for many practical purposes. Biological, agricultural, and other scientific experiments frequently deal with comparatively few observations. Sometimes the cost of obtaining additional observations is prohibitive; sometimes, indeed, it is impossible to obtain more data, as might be true in the case of meteorological records. In manufacturing inspection, too, small samples are of frequent occurrence.

Most of this theory has been developed and unified by R. A. Fisher, who has shown how to make more accurate estimates and how to utilize the maximum amount of information contained in a set of data, and has provided exact tests of reliability and significance. Fisher's efficient methods, at first slow in taking hold, because not thoroughly understood, gradually began to gain momentum, and are now spreading rapidly.

In this book I have endeavored to explain the most widely used of these methods, illustrating their application by comparatively simple numerical examples, so that the underlying principles are not lost sight of in a maze of arithmetical computations. The earlier chapters develop the fundamental concepts of statistics, so that the book is suitable as a textbook for a first course in the subject. It is also planned for those with some knowledge of the subject who wish to gain an insight into the more modern methods, as it leads from the classical concepts, through such topics as "Student's" distribution and the chi-square distribution, to the analysis of variance and the design of experiments, which are the culminating features of Fisher's work

Grateful acknowledgment is made to Professor Fisher, and to his publishers, Messrs. Oliver and Boyd, for their generous permission to reproduce, from "Statistical Methods for Research Workers," the tables of t, chi square, and the 5 per cent and 1 per cent points of the distribution of z.

I am deeply indebted to Dr. Churchill Eisenhart for a critical reading of the manuscript, and for many valuable suggestions. However, full responsibility for errors is, of course, my own.

I am also indebted to various persons and various sources for material used in the exercises, being particularly grateful to Messrs. A. G Brooks and J B. Gibson for supplying, and for granting permission to use, certain data of the Western Electric Company.

PAUL R. RIDER

WASHINGTON UNIVERSITY
SAINT LOUIS
September, 1938

CONTENTS

INTRODUCTION TO MODERN STATISTICAL METHODS

CHAPTER I

FREQUENCY DISTRIBUTIONS

1. Frequency tables. A *frequency table* is a table classifying a set of observations according to the numbers of them which fall within certain limits. It is a tabular method of exhibiting a *frequency distribution.* For example, Table 1 classifies the heights of a group of men. The table shows the frequency with which men of a given height, or rather between two given limits of height, occur in the group of 346 men under consideration. The values 58 inches, 60 inches, 62 inches, etc., are the *class limits,* and the difference between two consecutive class limits, here 2 inches, is the *class interval.* The mid-values of the classes are obviously 59 inches, 61 inches, etc. The *range* of the tablc, from 58 inches to 74 inches, is 16 inches.

TABLE 1

FREQUENCY TABLE OF HEIGHTS OF A GROUP OF MEN

Height in inches	Number of men within given limits of height (frequency)
58–60	1
60–62	2
62–64	9
64–66	48
66–68	131
68–70	102
70–72	40
72–74	13
Total	346

A word regarding classification and class limits may be worth while at this point. It is assumed that in the construction of Table 1 the measurements of height have been made to a sufficient degree of fineness that no doubt exists regarding the class to which a man belongs. If, however, we had a set of data in which measurements were made to the nearest inch, we could not employ the same class limits as those used in Table 1. For a height recorded as 62 inches would simply mean that the measurement was between 61.5 and 62.5 inches. In such a case, if we wished to use a 2-inch interval, we could set the classes as 57.5–59.5, 59.5–61.5, . . . The mid-values of these classes would be 58.5, 60 5, . . .

Some question arises as to what disposition to make of an observation or measurement which falls exactly on a class boundary. For example, if the classes are as in Table 1 and we have a measurement of 62 inches, it is not clear whether this should be assigned to the class 60–62 or to the class 62–64. In such an instance there are certain theoretical advantages in dividing the unit of frequency between the two classes, and assigning ½ to each class.

Difficulties in the classification of raw data can usually be avoided by a proper choice of class limits.*

2. Cumulative frequency tables. The above frequency distribution is equally well specified if we know the number of men below (or above) any given height, for if we know that there are 60 men below 66 inches in height and 12 below 64 inches, we can find at once that there are 60 − 12, or 48, men who are between 64 and 66 inches tall. If we convert Table 1 into a table showing the number of men below certain heights, we obtain the cumulative frequency table, Table 1A.

3. Continuous and discrete variables. The variable in the foregoing example is height. Theoretically it can be measured to any degree of fineness. Such a variable is called *continuous.* There are, however, variables which can have only integral values. Such variables are called *integral* or *discrete.* Examples of such variables are the number of petals on flowers (see Table 2), the number of spots obtained in throwing dice, the number of heads obtained in tossing coins, the number of children in a family.

* For a good discussion of these and related questions, see Yule and Kendall, "An Introduction to the Theory of Statistics," Charles Griffin & Co., Ltd., London, 1937.

TABLE 1A

CUMULATIVE FREQUENCY TABLE OF HEIGHTS

Height in inches	Number of men below specified height (cumulative frequency)
60	1
62	3
64	12
66	60
68	191
70	293
72	333
74	346

TABLE 2

FREQUENCY TABLE OF NUMBERS OF PETALS ON A CERTAIN SPECIES OF FLOWER

Number of petals	Number of flowers having a specified number of petals (frequency)
5	133
6	55
7	23
8	7
9	2
10	2
Total	222

4. Graphic representation of frequency distributions. In the case of continuous variables, the accepted method of representing a frequency distribution graphically is by means of a *rectangular frequency diagram* or *histogram.* This is constructed by marking

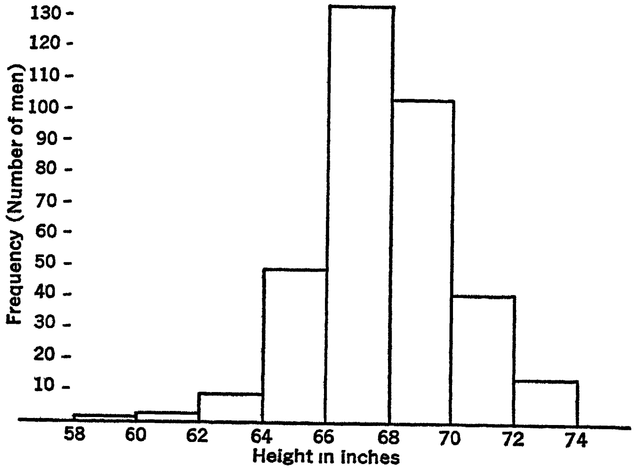


FIG. 1—Histogram of Heights of a Group of Men.

off a scale for the variable and erecting, at the appropriate positions on this scale, rectangles whose areas are equal or proportional to the respective class frequencies (See Fig 1.) In the usual case of class intervals of equal size, such rectangles will obviously have heights equal or proportional to the respective class frequencies.

A cumulative frequency distribution can be represented by plotting points with ordinates equal to the cumulated frequencies, and with abscissas equal to the upper limits of the classes, and then connecting these by straight-line segments. (See Fig. 3.) For discrete variables, perhaps the best method of graphic representation is to erect, at the proper places on the scale or base line, ordinates equal or proportional to the frequencies. (See Fig. 2.)

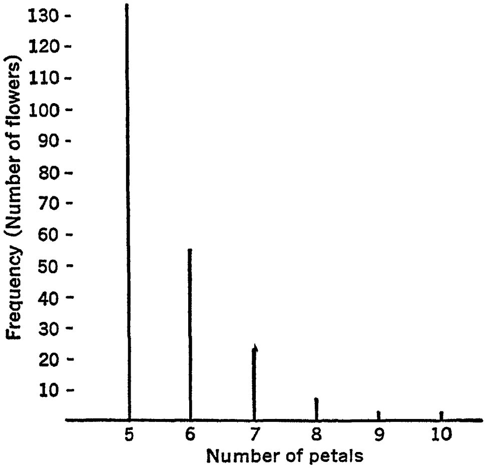


FIG 2—Frequency Diagram Discrete Variable—Number of Petals on a Species of Flower.

5. Frequency curves. Theoretical frequency distributions. If the size of the class interval of a distribution be decreased indefinitely and the number of individuals be simultaneously increased indefinitely, the histogram approaches a *frequency curve.* A frequency curve may be regarded as representing an idealized frequency distribution.

Suppose that the equation of the frequency curve is $Y = f(X)$.

The function $f(X)$ can always be multiplied by a constant factor which will make the area under the curve equal to unity, and we shall assume that this has been done. Then we have

$$\int_{-\infty}^{\infty} f(X)dX = 1 \tag{1}$$

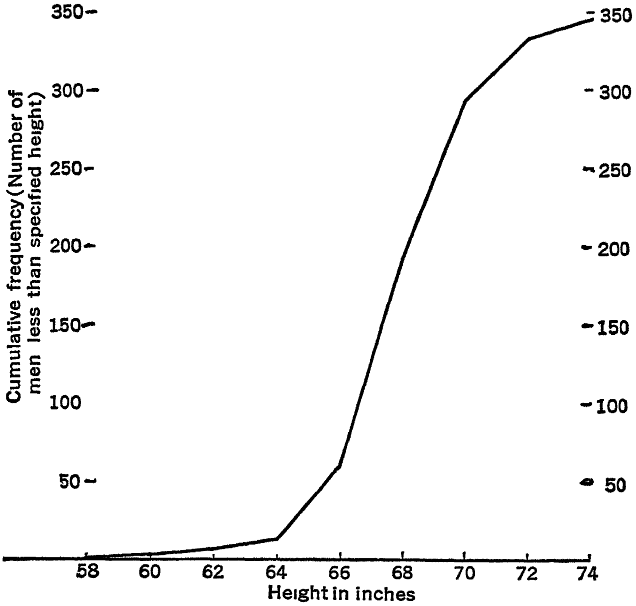


FIG 3—Cumulative Frequency Diagram.

(If the curve does not actually extend to infinity, but meets the X-axis in some point such as X_0 in Fig. 4, the value of $f(X)$ is to be regarded as zero outside of such points. Consequently we can always use $-\infty$ and ∞ as limits of the above definite integral.) The proportion of items between the values $X = a$ and $X = b$, $a < b$, is the shaded area in Fig. 4, and, provided (1) holds, is given analytically by

$$\int_{a}^{b} f(X)dX \quad . \tag{2}$$

By the artifice of defining $f(X) = 0$ outside the range of the curve, we see that we have defined $f(X)$ so that (2) gives the proportion of the area under the curve lying between *any* two values $X = a$ and $X = b$, $a < b$. This is the exact mathematical interpretation of $f(X)$. As an aid to intuition, it is often convenient to regard $f(X)dX$ as giving approximately the proportion of items in the interval between X and $X + dX$, the closeness of the approximation depending on how small dX is. It is desirable to write the equation of a frequency curve in the form $YdX = f(X)dX$, because if any transformation is made in the variable X it must also be made in the differential dX. (See next paragraph.)

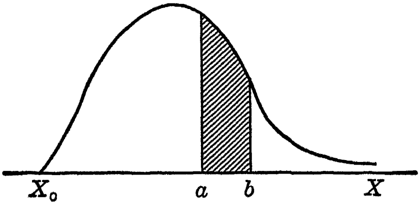


FIG. 4.—Frequency Curve.

The most widely used frequency curve is the *normal* curve, whose equation may be written

$$YdX = (2\pi)^{-\frac{1}{2}}e^{-(X-\mu)^2/2\sigma^2}\frac{dX}{\sigma} \tag{3}$$

If we make the transformation

$$\frac{X-\mu}{\sigma} = x, \quad \frac{dX}{\sigma} = dx$$

(3) goes into the simpler form

$$YdX = (2\pi)^{-\frac{1}{2}}e^{-x^2/2}dx \tag{4}$$

Another important curve is the *Pearson type III* curve

$$YdX = \frac{1}{k!}X^k e^{-X}dX, \quad 0 \leqq X < \infty \tag{5}$$

The symbol $k!$, called " k factorial," is defined by

$$k! = \int_0^\infty X^k e^{-X}dX = \Gamma(k+1) \tag{6}$$

If k is an integer, this reduces to $k(k-1)\ldots 3.2.1.$

Examples of ideal or theoretical discrete frequency distributions are the *binomial* distribution

$$Y = \frac{N!}{X!(N-X)!}p^X(1-p)^{N-X}, \quad 0 < p < 1, \qquad (7)$$

$$X = 0,1,2, \ \ldots, N$$

and the *Poisson exponential* distribution

$$Y = \frac{e^{-\mu}\mu^X}{X!}, \quad X = 0,1,2, \ldots \qquad (8)$$

EXERCISES

1. The frequency in a class of Table A is the number of students receiving grades between the limits indicated. (*a*) Draw a histogram for the data of this table (*b*) Construct a cumulative frequency table from the data. (*c*) Draw a cumulative frequency diagram

2. Table B gives the number of men of a certain group whose weights fall within specified limits. (*a*) Draw a histogram for this frequency table. (*b*) Construct a cumulative frequency table, and (*c*) draw a cumulative frequency diagram.

3. (*a*) Reduce Table C to a percentage frequency basis, and (*b*) draw the corresponding histogram. (*c*) Construct a cumulative percentage frequency table, and (*d*) draw the corresponding cumulative diagram

4. Table D shows the number of lost articles turned in per day at the lost and found bureau of a store The frequency is the number of days on which the specified number of articles were returned (*a*) Draw a frequency diagram (*b*) Construct a cumulative frequency table (*c*) Draw a cumulative frequency diagram.

5. (*a*) Draw a frequency diagram for the data of Table E (*b*) Construct a cumulative table, and (*c*) draw a cumulative diagram

6. Form a frequency table from the data of Table F. First construct a tally sheet making a mark for each time that a temperature of a given number of degrees occurs. From this tally sheet make a frequency table of class interval 1°. Then choose what you consider an appropriate wider class interval and group the frequencies accordingly. Next construct a rectangular frequency diagram. Finally construct a cumulative frequency table and the corresponding cumulative frequency diagram.

TABLE A

Grades Received by a Class of Students in an Examination

Grade	Frequency
10– 20	1
20– 30	2
30– 40	4
40– 50	6
50– 60	7
60– 70	12
70– 80	16
80– 90	10
90–100	4

TABLE B

Weights of a Group of Men

Weight in pounds	Frequency
100–110	2
110–120	3
120–130	11
130–140	34
140–150	84
150–160	65
160–170	48
170–180	33
180–190	20
190–200	11
200–210	4
210–220	3
220–230	1
230–240	1

TABLE C

Distribution of Employees in a Certain Industry According to Annual Earnings

Annual earnings, dollars	Number of employees
0– 200	88
200– 400	236
400– 600	396
600– 800	385
800–1000	412
1000–1200	341
1200–1400	208
1400–1600	113
1600–1800	68
1800–2000	68
2000–2200	33
2200–2400	18
2400–2600	15

TABLE D

Lost Articles Returned

Number of articles	Frequency
0	84
1	67
2	37
3	16
4	5
5	1

TABLE E

NUMBER OF CHILDREN BORN PER FAMILY IN 735 FAMILIES

Number of children born per family	Number of families
0	96
1	108
2	154
3	126
4	95
5	62
6	45
7	20
8	11
9	6
10	5
11	5
12	1
13	1

TABLE A

Grades Received by a Class of Students in an Examination

Grade	Frequency
10– 20	1
20– 30	2
30– 40	4
40– 50	6
50– 60	7
60– 70	12
70– 80	16
80– 90	10
90–100	4

TABLE B

Weights of a Group of Men

Weight in pounds	Frequency
100–110	2
110–120	3
120–130	11
130–140	34
140–150	84
150–160	65
160–170	48
170–180	33
180–190	20
190–200	11
200–210	4
210–220	3
220–230	1
230–240	1

TABLE C

Distribution of Employees in a Certain Industry According to Annual Earnings

Annual earnings, dollars	Number of employees
0– 200	88
200– 400	236
400– 600	396
600– 800	385
800–1000	412
1000–1200	341
1200–1400	208
1400–1600	113
1600–1800	68
1800–2000	68
2000–2200	33
2200–2400	18
2400–2600	15

TABLE D

Lost Articles Returned

Number of articles	Frequency
0	84
1	67
2	37
3	16
4	5
5	1

TABLE E

NUMBER OF CHILDREN BORN PER FAMILY IN 735 FAMILIES

Number of children born per family	Number of families
0	96
1	108
2	154
3	126
4	95
5	62
6	45
7	20
8	11
9	6
10	5
11	5
12	1
13	1

TABLE F

Hourly Temperatures in Degrees Fahrenheit, St. Louis, Mo , September, 1937
(Department of Agriculture, Weather Bureau)

Date	A. M.												P. M.												Date
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	
1	78	78	77	77	76	76	78	81	84	85	88	88	90	90	90	92	91	88	87	87	86	85	82	81	1
2	81	80	79	76	76	76	78	80	81	83	82	85	88	86	86	85	86	81	79	78	78	77	76	75	2
3	75	75	75	74	73	73	74	75	75	78	79	81	83	84	84	84	84	82	80	79	78	78	77	77	3
4	76	76	75	75	75	75	75	75	75	77	77	82	85	86	81	83	78	77	74	72	71	70	69	63	4
5	68	66	65	65	65	65	65	65	67	69	73	77	79	78	78	80	79	77	74	71	69	67	66	66	5
6	65	65	64	62	62	61	62	66	67	69	74	76	80	83	84	84	84	82	80	77	75	73	72	71	6
7	70	69	68	66	66	65	66	70	75	80	80	83	84	85	86	84	84	83	81	80	76	74	73	72	7
8	72	71	71	71	70	70	70	71	73	75	79	82	84	86	86	86	85	84	83	81	77	77	75	73	8
9	73	72	71	70	69	70	72	74	78	81	81	84	84	87	87	88	86	84	83	82	80	80	80	79	9
10	79	78	75	72	73	74	74	75	76	78	80	82	84	88	77	82	84	80	79	77	75	73	72	71	10
11	70	69	68	66	64	63	62	63	64	67	69	72	74	76	78	79	79	77	75	74	73	72	70	68	11
12	68	67	65	65	63	63	63	66	68	70	71	75	75	76	74	73	71	70	68	66	65	63	61	60	12
13	59	58	56	55	53	53	52	57	61	66	65	67	67	70	70	72	73	70	68	67	66	64	63	62	13
14	62	60	59	59	57	56	56	60	64	68	72	75	77	80	81	83	82	80	76	74	72	70	71	72	14
15	72	72	70	68	66	65	65	68	71	71	72	73	74	76	77	78	78	77	76	73	70	67	65	62	15
16	60	56	55	53	51	50	51	54	56	57	59	61	63	65	65	66	65	64	63	61	59	58	57	56	16
17	54	53	52	52	52	50	49	56	60	62	64	66	67	69	71	69	69	68	66	64	61	59	57	57	17
18	57	54	53	54	54	52	52	56	60	66	69	76	77	79	80	82	81	78	76	73	72	71	70	70	18
19	70	70	70	69	65	62	60	60	63	66	69	72	74	74	76	75	75	72	69	66	63	60	59	58	19
20	57	56	54	52	51	50	53	59	64	68	69	72	73	76	76	77	76	74	73	71	68	66	65	63	20
21	61	60	60	60	59	58	57	63	69	76	81	83	85	88	88	87	85	83	81	79	77	74	73	71	21
22	71	69	67	66	65	66	67	72	77	81	85	88	89	91	92	92	91	87	85	82	81	79	76	76	22
23	75	74	73	72	71	72	73	77	80	82	85	87	88	88	89	87	86	83	82	79	78	77	76	76	23
24	76	74	74	73	73	73	73	74	78	82	82	78	78	80	82	73	71	71	66	62	61	60	59	59	24
25	59	59	59	58	58	56	55	55	56	55	58	61	61	63	65	65	63	61	59	57	56	55	54	52	25
26	52	51	50	50	50	48	48	50	52	57	59	60	61	61	61	61	62	61	60	56	54	54	53	51	26
27	50	50	48	48	47	47	48	51	56	59	60	63	63	62	63	63	62	60	60	59	57	56	54	53	27
28	51	50	49	49	49	49	49	51	58	62	64	67	66	67	70	71	71	67	64	63	61	60	59	59	28
29	59	59	60	58	57	57	58	63	67	69	72	76	78	81	80	79	78	74	73	70	67	66	65	64	29
30	64	63	62	61	62	62	61	69	73	76	78	82	84	86	86	86	85	81	78	74	73	71	69	68	30

CHAPTER II

AVERAGES AND MOMENTS

6. Averages. An *average* is a quantity which may be regarded as representative of a group of data. We may use an average to characterize a frequency distribution, or we may use the averages of two different distributions for purposes of comparison. For example, we may compare the average weights of two football teams, or the average wages in two different occupations.

There are various types of average, such as the arithmetic mean, the geometric mean, the harmonic mean, the median, and the mode.

7. Arithmetic mean. The *arithmetic mean,* often called merely the *mean,* of a set of quantities is their total divided by their number. Thus, if we have N quantities $X_1, X_2, \ldots, X_N$, their mean is

$$\bar{X} = \frac{1}{N} \Sigma X = \frac{1}{N} (X_1 + X_2 + \ldots + X_N) \qquad (1)$$

If each X_i occurs with the respective frequency f_i, the mean may be written as

$$\bar{X} = \frac{\Sigma Xf}{\Sigma f} = \frac{1}{N} (X_1 f_1 + X_2 f_2 + \ldots + X_k f_k) \qquad (2)$$

where $N = \Sigma f = f_1 + f_2 + \ldots + f_k$. However, equation (1) will mean the same thing as equation (2) if we consider that some of the X's in (1) may be identical in value

The arithmetic mean has the property that the algebraic sum of deviations from it is zero. That is,

$$\Sigma(X - \bar{X}) = 0 \qquad (3)$$

In computing the mean, it is sometimes convenient to choose an arbitrary origin $\bar{X}'$. Then the mean may be found by the formula

$$\bar{X} = \bar{X}' + \frac{1}{N}\Sigma(X - \bar{X}') \tag{4}$$

For example, if we wish to find the mean of the numbers 205, 197, 200, 204, we might choose the origin 200, the deviations from which are 5, −3, 0, 4, respectively. Then

$$\bar{X} = 200 + \tfrac{1}{4}(5 - 3 + 0 + 4) = 201.5$$

For application to a frequency table, formula (4) might be written in the form

$$\bar{X} = \bar{X}' + \frac{\Sigma x'f}{\Sigma f}c \tag{5}$$

in which x' is the deviation, in terms of the class interval c as a unit.

Let us apply this formula to the computation of the arithmetic mean of the heights in Table 1. The details of the computation are shown in Table 3, in which the heights are regarded as concentrated at the mid-values of the classes.

TABLE 3

Computation of the Mean Height of a Group of Men

Height in inches X	Deviation in class intervals from 67 x'	Frequency f	$x'f$
59	−4	1	− 4
61	−3	2	− 6
63	−2	9	−18
65	−1	48	−48
67	0	131	0
69	1	102	102
71	2	40	80
73	3	13	39
Total		346	145

Here $\bar{X}' = 67$ inches, $c = 2$ inches, and we find

$$\bar{X} = 67 + \tfrac{145}{346} \times 2 = 67.84 \text{ in.}$$

The arithmetic mean of a continuous distribution represented by the frequency curve $Y = f(X)$ is given by the formula

$$\bar{X} = \frac{\int X f(X)dX}{\int f(X)dX} \tag{6}$$

or simply by the numerator if the area under the curve is taken as unity.

8. Weighted mean. In the computation of a mean the various quantities may be assigned various *weights*. If w_i is the weight associated with X_i then

$$\text{Weighted mean} = \frac{\Sigma w_i X_i}{\Sigma w_i} \tag{7}$$

The weights may be somewhat arbitrary. For example, in computing a student's final grade the instructor might assign various weights to the laboratory grade, the final examination grade, the test average, and the average of daily recitations, depending upon the importance that he attaches to the various phases of the work. On the other hand, they may be somewhat more definite; a student's average in all his work would in general be found by weighting the grade in each course by the number of hours per week that the course meets. Comparison of (2) and (7) shows that the mean of a frequency table may be regarded as a weighted mean in which the weights are the frequencies.

9. Median. The *median* of a set of quantities is the middle value when the quantities are arranged in order of magnitude. Thus, the median of 1, 3, 8, 10, 20 is 8. Obviously when there is an even number of quantities there is no middle quantity; in such a case the median is usually defined as the number halfway between the middle pair, that is, the arithmetic mean of the middle pair. The median of 1, 3, 8, 10, 20, 25 would be $\frac{1}{2}(8 + 10) = 9$.

The serial number of the median of N quantities is $(N + 1)/2$. For 99 quantities the serial number is $(99 + 1)/2 = 50$, and the median is the fiftieth quantity when they are arranged in order. There will be 49 quantities below the median and 49 above. For 100 quantities, $(N + 1)/2 = 101/2 = 50.5$, and the median is halfway between the fiftieth and the fifty-first quantity.

To find the median in a frequency table, that is, in a grouped frequency distribution, we must resort to interpolation. A satisfactory interpolation formula, based upon the assumption that the quantities are uniformly distributed in the class interval in which the median lies, is

$$\text{Median} = l + \frac{\frac{1}{2}(N + 1) - N_b - \frac{1}{2}}{N_m} c \tag{8}$$

which reduces to the simpler form

$$\text{Median} = l + \frac{\frac{1}{2}N - N_b}{N_m} c \tag{9}$$

in which l = lower limit of median class.
N = total frequency.
N_m = frequency of median class.
N_b = sum of frequencies below median class.
c = class interval.

Let us apply this formula and find the median height of the group of men in Table 1. Here $N = 346$, and the serial number of the median is $(346 + 1)/2 = 173.5$. Referring to the cumulative frequency table, Table 1A, we see that there are 60 men below 66 inches in height, and 191 below 68 inches, so that the median is in the class 66–68 inches, that is, the class whose lower limit is 66 inches. Employing formula (9), we have

$$\begin{aligned}\text{Median} &= l + \frac{\frac{1}{2}N - N_b}{N_m} c \\ &= 66 + \frac{173 - 60}{131} \times 2 = 67.73 \text{ in.}\end{aligned}$$

For a continuous distribution the median is the value of that abscissa corresponding to the ordinate which divides the area under the frequency curve into two equal parts. Analytically, if M is the median, then

$$\int_{-\infty}^{M} f(X)dX = \int_{M}^{\infty} f(X)dX \tag{10}$$

10. Mode. The *mode* is the value which has the greatest frequency. In a grouped frequency distribution we shall simply refer to the *modal class* (the class 66–68 inches in Table 1, for example), as a precise determination of the mode is difficult; perhaps the only satisfactory way is to fit a theoretical frequency curve to the data and then to find the abscissa corresponding to its maximum point.

11. Geometric mean. The *geometric mean* of N quantities is the Nth root of their product. It is best found by means of logarithms. If G is the geometric mean of $X_1, X_2, \ldots, X_N$, then

$$G = (X_1X_2 \ldots X_N)^{1/N} = \left(\prod_{i=1}^{N} X_i\right)^{1/N} \quad \text{or} \quad (\Pi X)^{1/N} \tag{11}$$

$$\log G = \frac{1}{N}(\log X_1 + \ldots + \log X_N) = \frac{1}{N}\Sigma \log X \tag{12}$$

The geometric mean is useful in the construction of index numbers.

12. Harmonic mean. Suppose that an automobile makes a 200-mile trip, covering the first 100 miles at the rate of 50 miles an hour and the second 100 miles at the rate of 40 miles an hour. We can *not* find its average speed by taking ½(50 + 40) = 45 miles per hour. The total time is 2 hours for the first 100 miles, plus 2½ hours for the second 100 miles, or a total of 4½ hours. The average speed is therefore 200 ÷ 4½ = $44\frac{4}{9}$ miles per hour. The same result can be obtained by employing the *harmonic mean*, which is the reciprocal of the arithmetic mean of the reciprocals of a set of quantities. If H is the harmonic mean of $X_1, \ldots, X_N$, then

$$\frac{1}{H} = \frac{1}{N}\Sigma\frac{1}{X} \tag{13}$$

Applying this formula to the above example, we get

$$\frac{1}{H} = \frac{1}{2}\left(\frac{1}{50} + \frac{1}{40}\right) = \frac{9}{400}, \quad H = 44\tfrac{4}{9} = 44.\dot{4}$$

(A dot above a number following a decimal point means that the number is to be repeated indefinitely Thus 1 8$\dot{3}$ means 1.8333 . . . . Similarly 1.$\dot{2}1\dot{6}$ means 1 216216216 . . . .)

13. Appropriateness of different averages. Different averages may be used for different purposes For example, in economic statistics it is often desirable to disregard extreme items, which may be due to unusual circumstances. In such cases the median has been found to be serviceable, as it is not affected by extreme items.

It was stated above that the geometric mean is useful in the construction of index numbers.

In section 12 was given an example in which the harmonic mean was found to be the appropriate one to use.

It may be worth mentioning at this point that the arithmetic, geometric, and harmonic means of a set of quantities are always in the same order of magnitude. If we designate them by A, G, and H respectively, then

$$A \geqq G \geqq H \tag{14}$$

the equality signs holding only if all the quantites have the same value. For example, if we have the numbers 2, 4, 8, we find

$$A = \frac{1}{3}(2 + 4 + 8) = \frac{14}{3} = 4\tfrac{2}{3},$$

$$G = (2 \times 4 \times 8)^{1/3} = (64)^{1/3} = 4$$

$$\frac{1}{H} = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) = \frac{1}{3} \times \frac{7}{8} = \frac{7}{24},$$

$$H = \frac{24}{7} = 3\tfrac{3}{7}$$

and we see that the values are arranged in the order of magnitude specified by (14).

If a man were purchasing a house and wanted to know the average time that it would take him to reach his place of business he would probably find the mode the most acceptable average to employ, since he would doubtless like to know how long it would *usually* take him to make the trip.

Although the various averages other than the arithmetic mean arc thus secn to have their merits, in certain cases, nevertheless, this familiar average has advantages which in most circumstances outweigh those of the others Chief among these is its reliability in sampling. One of the important uses of an average calculated from a sample is that of estimating the corresponding average in the population, that is, the large group from which the sample is drawn. Suppose that the arithmetic mean of a sample is 10 and we estimate that the population mean is between 8 and 12; suppose, on the other hand, that the median is also 10 and we estimate that the population median is between 8 and 12. We are more likely to be correct in our first estimate than in our second, unless we are sampling from an unusual population.

14. Partition values. Quite analogous to the median are the *quartiles*, the *deciles*, and the *percentiles*. These are not averages, but might be termed *partition values*. The quartiles divide the frequencies into four equal groups, the deciles into ten, and the percentiles into one hundred. The serial number of the first or lower quartile is $(1/4)(N + 1)$, that of the third or upper quartile is $(3/4)(N + 1)$. The second quartile is the median. The serial number of the kth decile is $(k/10)(N + 1)$, and that of the kth percentile is $(k/100)(N + 1)$.

Formula (8) of section 9 may be generalized to give any partition value. Thus, the partition which has a proportion p of items below it is given by the formula

$$l + \frac{p(N + 1) - N_b - \frac{1}{2}}{N_p} c$$

the meanings of the symbols of which will be obvious if section 9 is reread.

To illustrate the application of this formula, let us compute the third quartile, Q_3, of the distribution of heights in Table 1. We find the serial number of the third quartile to be $(3/4)(346 + 1) = 260.25$. From Table 1A we see that Q_3 is in the class whose lower limit is 68 inches. Thus,

$$Q_3 = 68 + \frac{(\frac{3}{4})347 - 191 - \frac{1}{2}}{102} \times 2 = 69.348 \text{ in}$$

15. Variance and standard deviation. The *variance* of a set of quantities is defined as the sum of the squares of their deviations from their mean, divided by their number, that is, their mean square deviation from their mean. Analytically the variance is defined by

$$\sigma^2 = \frac{1}{N}\Sigma(X - \bar{X})^2 \quad \text{or} \quad \frac{1}{N}\Sigma\,(X - \bar{X})^2 f \tag{15}$$

for a discrete distribution, and by

$$\sigma^2 = \int (X - \bar{X})^2 f(X)dX \div \int f(X)dX \tag{16}$$

$$= \int (X - \bar{X})^2 f(X)dX \quad \text{if} \quad \int f(X)dX = 1 \tag{17}$$

for a continuous distribution. If $x = X - \bar{X}$, (15) and (17) assume respectively the simpler forms

$$\sigma^2 = \frac{1}{N}\Sigma x^2 \quad \text{or} \quad \frac{1}{N}\Sigma x^2 f \tag{18}$$

$$\sigma^2 = \int x^2 f(x)dx \quad \text{if} \quad \int f(x)dx = 1 \tag{19}$$

The *standard deviation*, σ, is the square root of the variance; that is, it is the root-mean-square deviation about the mean.

It can be shown that *the sum of the squares of the deviations of a set of quantities from any fixed value is a minimum when that fixed value is the mean.* In other words, the variance is the minimum mean square deviation and the standard deviation is the minimum root-mean-square deviation.

For computational purposes the formula

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{\Sigma X^2}{N} - \bar{X}^2 \tag{20}$$

is better than (15), from which it can easily be derived. In (20), X may be measured from any origin whatever. It is often convenient and simpler to choose an origin near the mean.

The method of computing the variance from a grouped frequency distribution will be illustrated in Table 4. Here the data of Table 1 have been used.

TABLE 4

COMPUTATION OF THE VARIANCE OF THE HEIGHTS OF A GROUP OF MEN

Height in inches X	Deviation from 67 x'	Frequency f	$x'f$	x'^2f	$(x'+1)^2f$
59	−4	1	− 4	16	9
61	−3	2	− 6	18	8
63	−2	9	−18	36	9
65	−1	48	−48	48	0
67	0	131	0	0	131
69	1	102	102	102	408
71	2	40	80	160	360
73	3	13	39	117	208
Total		346	145	497	1133

The appropriate formula is

$$\sigma^2 = \left[\frac{\Sigma x'^2 f}{\Sigma f} - \left(\frac{\Sigma x' f}{\Sigma f}\right)^2\right] c^2 \qquad (21)$$

in which x' is the deviation (expressed in terms of the class interval) from any origin, preferably the center of a class in or near which the mean lies, and c is the class interval. From Table 4 we find

$$\sigma^2 = \left[\frac{497}{346} - \left(\frac{145}{346}\right)^2\right] \times 2^2 = 5.0432 \text{ in }^2$$

The final column of Table 4 affords a check, known as the *Charlier check*, on the accuracy of the totals of that table, since $\Sigma(x'+1)^2 f = \Sigma x'^2 f + 2\Sigma x' f + \Sigma f$, or $1133 = 497 + 290 + 346$.

The standard deviation is

$$\sigma = 1.123 \text{ class intervals} = 2.246 \text{ in.}$$

Incidentally we can find the mean (cf. section 7),

$$\bar{X} = \bar{X}' + \frac{\Sigma x' f}{\Sigma f} c = 67 + \frac{145}{346} \times 2 = 67.84 \text{ in.}$$

When a continuous variable is grouped into classes, an adjustment called *Sheppard's correction* is sometimes applied to the variance if the frequency distribution tapers off gradually at both ends. This correction is $-c^2/12$. If ${\sigma^*}^2$ represents the variance after Sheppard's correction is applied, then

$$ {\sigma^*}^2 = \sigma^2 - \frac{c^2}{12} \tag{22} $$

From this value we may, of course, obtain the corresponding value σ^* of the standard deviation. In the example given,

$$ {\sigma^*}^2 = 5.0432 - \frac{2^2}{12} = 4.7099 \text{ in.}^2, \quad \sigma^* = 2.17 \text{ in.} $$

The variance and the standard deviation are measures of the dispersion of a set of quantities or of a frequency distribution. A distribution which is widely scattered will have a larger variance and consequently a larger standard deviation than a more compact group.

It is sometimes desirable to compare the dispersion of two distributions. For example, although we should expect a larger actual dispersion in the weights of rabbits than in the weights of mice, the variability in relation to size might not be so great. This suggests dividing each standard deviation by the respective mean. The quotient, $\sigma/\bar{X}$ (usually expressed as a percentage), is called the *coefficient of variation.* It is independent of the unit of measurement and thus renders comparable not only the variability of two distributions such as the weights of rabbits and the weights of mice, but also the variability of two different characteristics such as weight and stature.

16. Mean deviation. Another measure of dispersion is the *mean deviation,* which is the mean of the absolute values of the deviations from any value A. Analytically, it is defined by

$$ \frac{1}{N} \Sigma \mid X - A \mid \tag{23} $$

for a discrete variable, and by

$$ \int \mid X - A \mid f(X) dX \quad \text{if} \quad \int f(X) dX = 1 \tag{24} $$

for a continuous variable. As *the mean deviation from the median is the minimum mean deviation,* A is usually set equal to the median.

For computing the mean deviation of a grouped frequency distribution the following formula* is recommended:

$$\text{Mean deviation} = \frac{c}{N}[\Sigma\,|\,x'\,|\,f + (N_b - N_a)d + N_m(d^2 + \tfrac{1}{4})] \quad (25)$$

Here x' = deviation, in class-interval units, from center of class containing average from which deviations are measured.

d = distance, in class-interval units, from center of this class to average.

N_b = total frequency below this class.

N_a = total frequency above this class.

N_m = frequency of this class.

N = total frequency.

c = class interval.

17. Moments. The mean and the variance are special cases of moments, the kth *moment* of $X_1, \ldots, X_N$ being defined as

$$\mu'_k = \frac{1}{N}\,\Sigma X^k \qquad (26)$$

The kth moment of a continuous distribution is

$$\mu'_k = \int X^k f(X)dX \quad \text{if} \quad \int f(X)dX = 1 \qquad (27)$$

The kth moment about the mean is

$$\mu_k = \frac{1}{N}\,\Sigma x^k \quad \text{or} \quad \mu_k = \int x^k f(x)dx, \quad x = X - \bar{X} \qquad (28)$$

Obviously μ'_1 is the mean and μ_2 the variance, that is, $\mu'_1 = \bar{X}$, $\mu_2 = \sigma^2$. (The word "moment" will ordinarily be understood to signify moment about the mean.)

We shall illustrate the method of transferring from moments

* For its derivation and an illustration of its application the student is referred to H L. Rietz (Editor), "Handbook of Mathematical Statistics," Houghton Mifflin Co , Boston, 1924, pp 29–31

about any origin to moments about the mean, by developing the formula for the third moment.

$$\begin{aligned}\mu_3 &= \frac{1}{N}\,\Sigma x^3 = \frac{1}{N}\,\Sigma(X - \bar{X})^3 \\ &= \frac{1}{N}\,(\Sigma X^3 - 3\bar{X}\Sigma X^2 + 3\bar{X}^2\Sigma X - N\bar{X}^3) \\ &= \mu_3' - 3\bar{X}\mu_2' + 2\bar{X}^3 \\ &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3\end{aligned}$$

The formulas for the first four moments are as follows:

$$\begin{aligned}&\mu_1 = 0, \quad \mu_2 = \mu_2' - \mu_1'^2 \\ &\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \\ &\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4\end{aligned} \tag{29}$$

Others can be developed as shown above. All these formulas hold for continuous as well as discrete variables. It will be noted that the formula for μ_2 is the same as (20).

The adjusted moments, after *Sheppard's corrections* for grouping a continuous variable (see section 15) have been applied, are

$$\begin{aligned}&\overset{*}{\mu}_1 = \mu_1 = 0, \quad \overset{*}{\mu}_2 = \mu_2 - \frac{c^2}{12}, \\ &\overset{*}{\mu}_3 = \mu_3, \quad \overset{*}{\mu}_4 = \mu_4 - \frac{\mu_2}{2}\,c^2 + \frac{7}{240}\,c^4\end{aligned} \tag{30}$$

in which c is the class interval employed in the grouping.

The following functions of moments are sometimes used:

$$\alpha_3 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \text{ (often denoted by } \sqrt{\beta_1}) \tag{31}$$

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \text{ (often denoted by } \beta_2) \tag{32}$$

The quantity α_3 is a measure of the *skewness* or lack of symmetry of a frequency distribution. For a curve which has a longer tail to the right, such as the curve in Fig. 4, the skewness is positive;

when the curve stretches out more to the left the skewness is negative.

For the normal distribution α_4 has the value 3, for curves which come to a sharper peak than the normal curve it has a value greater than 3, while for curves that are flatter than the normal curve it has a value less than 3. The quantity $\alpha_4 - 3$, the *excess* of the value of α_4 over the value for the normal distribution, is a measure of the *kurtosis* of the distribution.

The computation of the first four moments of the frequency distribution of Table 1 is shown in Table 5.

TABLE 5

Computation of the First Four Moments of the Frequency Distribution of Heights of a Group of Men

Height in inches	x'	f	$x'f$	x'^2f	x'^3f	x'^4f	$(x'-1)^4f$
59	−4	1	− 4	16	−64	256	625
61	−3	2	− 6	18	−54	162	512
63	−2	9	−18	36	−72	144	729
65	−1	48	−48	48	−48	48	768
67	0	131	0	0	0	0	131
69	1	102	102	102	102	102	0
71	2	40	80	160	320	640	40
73	3	13	39	117	351	1053	208
Total		346	145	497	535	2405	3013

To verify the totals of Table 5 we may employ the Charlier check

$$\Sigma(x' \pm 1)^4 f = \Sigma x'^4 f \pm 4\Sigma x'^3 f + 6\Sigma x'^2 f \pm 4\Sigma x' f + \Sigma f \qquad (33)$$

Here it seems more advisable to use the lower signs. We find

$$3013 = 2405 - 4 \times 535 + 6 \times 497 - 4 \times 145 + 346$$

We now calculate

$$\mu_1' = \tfrac{145}{346}\, c = 0.41908\, c, \qquad \mu_2' = \tfrac{497}{346}\, c^2 = 1.43642\, c^2$$

$$\mu_3' = \tfrac{535}{346}\, c^3 = 1.54624\, c^3, \qquad \mu_4' = \tfrac{2405}{346}\, c^4 = 6.95087\, c^4$$

$\mu_2 = 1.43642\ c^2 - (0.41908\ c)^2 = 1.26079\ c^2 = 5.04316$ in.^{2}

$\mu_3 = 1.54624\ c^3 - 3(1.43642\ c^2)(0.41908\ c) + 2(0.41908\ c)^3 = -0.11247\ c^3$
$= -0.89976$ in.^{3}

$\mu_4 = 6.95087\ c^4 - 4(1.54624\ c^3)(0.41908\ c) + 6(1.43642\ c^2)(0.41908\ c)^2 - 3(0.41908\ c)^4$
$= 5.77997\ c^4 = 92.47952$ in.^{4}

$\mu_2^* = (1.26079 - \frac{1}{12})\ c^2 = 1.17746\ c^2 = 4.70984$ in.^{2}

$\mu_4^* = (5.77997 - \frac{1}{2} \times 1.26079 + \frac{7}{240})\ c^4 = 5.17874\ c^4 = 82.85984$ in.^{4}

EXERCISES

1. The following observations were made on the vertical diameter of the planet Venus (the unit is 1 second of angle): 42.70, 42.56, 43.01, 43.48, 42.76, 43.06, 43.63, 42.87, 41.60, 42.78, 42.95, 43.20, 43.18, 43.39, 43.10. Find (*a*) the arithmetic mean, (*b*) the median, (*c*) the standard deviation, (*d*) the mean deviation from the mean, (*e*) the mean deviation from the median.

In the next five exercises find, for the tables mentioned, the following quantities: (*a*) arithmetic mean, (*b*) median, (*c*) upper and lower quartiles, (*d*) mode, (*e*) standard deviation, (*f*) mean deviation from median, (*g*) third moment, (*h*) fourth moment.

2. Table A, page 8 **3.** Table B, page 8.
4. Table C, page 8. **5.** Table D, page 8.
6. Table E, page 9.

7. From the results of exercise 6, page 7, calculate the values of the following quantities from the raw data, and also from the grouped data, and compare: (*a*) mean, (*b*) median, (*c*) standard deviation, (*d*) third moment, (*e*) fourth moment.

8. The relative price of a commodity is the ratio, usually expressed as a percentage, of its price at a given time to its price during a specified base period. Find the geometric mean of the following set of relative prices: 142, 156, 94, 175, 150, 114, 150, 95, 72, 119.

9. Given N pairs of numbers $(X_1', X_1''), \ldots, (X_N', X_N'')$. Let G' be the geometric mean of the quantities $X_1', \ldots, X_N'$, G'' the geometric mean of $X_1'', \ldots, X_N''$, and G the geometric mean of the ratios $X_1'/X_1'', \ldots, X_N'/X_N''$. Show that $G = G'/G''$.

10. A man motors from A to B. A large part of the distance is uphill, and he gets a mileage of only 10 miles per gallon of gasoline. On the return trip he makes 15 miles per gallon. Find the harmonic mean of his mileage. Verify the fact that this is the proper average to use by assuming that the distance from A to B is 60 miles.

11. In five different cities the numbers of streetcar tickets sold for a dollar are respectively: 8, 12, 10, 8, 9. Find the harmonic mean of the number of tickets sold for a dollar.

12. Table G gives index numbers for various items entering the cost of living. Find an index of the cost of living by computing a weighted average of these items. The weights to be used are also given in the table.

TABLE G

	Index	Weight
Clothing	77 3	13
Food	74 5	43
Fuel and light	85 8	6
Housing	64 6	18
Sundries	92 5	20

13. Table H is the grade sheet of a university graduate (*a*) Find his average grade by weighting each course taken by the number of units of credits which the course carries (*b*) Find his average grade in this manner foi each of the four years. (*c*) Verify the statement that his average grade for the entire course is equal to the weighted average of his average grades for freshman, sophomore, junior, and senior years, the weights being the total numbers of credits for these respective years.

TABLE H

Grade Sheet of a University Student

	1st Semester			2nd Semester		
	Course number	Credits	Grade	Course number	Credits	Grade
Freshman Year	1	3	72	6	3	70
	2	5	79	7	5	80
	3	5	85	8	4	82
	4	3	72	9	3	89
	5	3	64	10	3	75
Sophomore Year	11	2	71	17	2	76
	12	5	86	18	3	78
	13	5	77	19	5	88
	14	2	82	20	3	72
	15	2	77	21	1	74
	16	1	73	22	4	63
Junior Year	23	3	60	33	3	75
	24	1	70	34	1	76
	25	.1	85	35	2	86
	26	2	96	36	2	97
	27	1	90	37	1	91
	28	3	72	38	3	60
	29	1	87	39	1	81
	30	3	85	40	3	78
	31	3	74	41	3	86
	32	3	85	42	3	78
Senior Year	43	3	78	51	3	74
	44	2	87	52	2	81
	45	2	70	53	2	72
	46	3	69	54	2	70
	47	1	82	55	1	84
	48	1	86	56	2	71
	49	2	83	57	1	85
	50	3	72	58	3	65

CHAPTER III

REGRESSION

18. Regression or trend lines. If the pairs of numbers $(X, Y) = (0,1), (1,3), (3,2), (6,5), (8,4)$ are plotted, as in Fig. 5,* we see that they tend to lie on a straight line. A line drawn so as to pass near most of the points is called a *regression line* if X and Y represent characteristics such as height and weight, for example, or a *trend line* if X represents time and Y represents some such quantity as population or the price of a commodity. The object of such a line is usually to estimate one of the variables, say Y, from the other, X.

A standard procedure in fitting a line to the points is the *method of least squares.* By this method we write the equation of a line

$$Y' = a + bX \tag{1}$$

and then determine a and b so that the sum of the squares of the vertical deviations of the points from this line will be the least possible. That is, we minimize $\Sigma(Y - Y')^2$. We could just as well minimize the sum of squares of the horizontal distances of the points from the line (or of their perpendicular distances), but this would in general give a different line. For a line of the form (1), X is termed the *independent* and Y the *dependent* variable.

If we have N pairs of values $(X_1, Y_1), \ldots, (X_N, Y_N)$ and substitute each X in (1) we obtain N values of Y'. We then wish to minimize

$$\Sigma(Y - Y')^2 = \Sigma(Y - a - bX)^2$$

Employing the usual methods of the calculus, we differentiate

* P. 30. (Such a figure is sometimes called a *scatter diagram*).

partially with respect to a and then b and set the derivatives equal to zero, obtaining the *normal equations*

$$\begin{aligned} aN + b\Sigma X &= \Sigma Y \\ a\Sigma X + b\Sigma X^2 &= \Sigma XY \end{aligned} \tag{2}$$

If we write

$$D = \begin{vmatrix} N & \Sigma X \\ \Sigma X & \Sigma X^2 \end{vmatrix} = N\Sigma X^2 - (\Sigma X)^2 \tag{3}$$

the normal equations have the solutions

$$a = \frac{1}{D}\begin{vmatrix} \Sigma Y & \Sigma X \\ \Sigma XY & \Sigma X^2 \end{vmatrix} = \frac{(\Sigma X^2)(\Sigma Y) - (\Sigma XY)(\Sigma X)}{N\Sigma X^2 - (\Sigma X)^2} \tag{4}$$

$$b = \frac{1}{D}\begin{vmatrix} N & \Sigma Y \\ \Sigma X & \Sigma XY \end{vmatrix} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \tag{5}$$

The quantity b is called the *coefficient of regression of* Y *on* X. It measures the average increase of Y per unit increase of X. If we determine a line of the form $X' = a + bY$, b is called the *coefficient of regression of* X *on* Y. The values of a and b in such an equation can obviously be obtained from (4) and (5) respectively by interchanging X and Y.

If the variables are measured from their respective means the foregoing equations are materially simplified. Let x and y represent deviations from the means of X and Y respectively. Then $\Sigma x = \Sigma y = 0$, and the normal equations reduce to $aN = 0$ and $b\Sigma x^2 = \Sigma xy$, with the solutions

$$a = 0, \quad b = \frac{\Sigma xy}{\Sigma x^2} \tag{6}$$

Even if only X is measured from its mean we have simpler results, viz.,

$$a = \frac{\Sigma Y}{N} = \overline{Y}, \quad b = \frac{\Sigma xY}{\Sigma x^2} \tag{7}$$

In any case the equation of the line of regression can be written

$$Y' - \overline{Y} = b(X - \overline{X}) \quad \text{or} \quad y' = bx \tag{8}$$

in which b may be obtained from (5), (6), or (7).

The mean product of deviations of X and Y from their respective means is called their *covariance.* It may be written in any of the following forms:

$$\frac{1}{N}\Sigma xy = \frac{1}{N}\Sigma(X - \bar{X})(Y - \bar{Y}) = \frac{1}{N}\Sigma XY - \bar{X}\bar{Y}$$
$$= \frac{1}{N}\Sigma xY = \frac{1}{N}\Sigma Xy \qquad (9)$$

If we divide numerator and denominator of b in (6) by N, we see that the coefficient of regression of Y on X is the covariance of X and Y divided by the variance of X, and coefficient of regression of X on Y is the covariance of X and Y divided by the variance of Y.

It is useful to know the sum of squares of deviations from the line of least squares. This can be found directly by substituting each X_i in the equation of the regression line, computing the corresponding Y_i', subtracting it from Y_i, squaring the result, and finally summing A shorter method is to make use of the formula

$$\Sigma(Y - Y')^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY \qquad (10)$$

which may be developed as follows:

$$\Sigma(Y - Y')^2 = \Sigma(Y - a - bX)^2$$
$$= \Sigma Y(Y - a - bX) - a\Sigma(Y - a - bX)$$
$$- b\Sigma X(Y - a - bX)$$

The last two terms drop out, since the normal equations (2) are satisfied, and the remaining term reduces to the right side of (10).

This sum of squares may be placed in another convenient form. If we replace a and b by their determinant values from (4) and (5), we find that

$$\Sigma(Y - Y')^2 = \begin{vmatrix} \Sigma Y^2 & \Sigma Y & \Sigma XY \\ \Sigma Y & N & \Sigma X \\ \Sigma XY & \Sigma X & \Sigma X^2 \end{vmatrix} \div \begin{vmatrix} N & \Sigma X \\ \Sigma X & \Sigma X^2 \end{vmatrix} \qquad (11)$$

It is to be noted that the denominator is the determinant D and that the numerator is this determinant bordered by $\Sigma Y^2, \Sigma Y, \Sigma XY$.

If we measure both variables from their means we can reduce (11) to the simpler form

$$\Sigma(y - y')^2 = \Sigma(Y - Y')^2 = \frac{1}{\Sigma x^2}\begin{vmatrix} \Sigma y^2 & \Sigma xy \\ \Sigma xy & \Sigma x^2 \end{vmatrix} \qquad (12)$$

To illustrate the foregoing theory we shall fit a line of least squares to the set of points given at the beginning of this section. (We should, of course, usually have more pairs of values than five, but the regression line could be obtained by the process illustrated) The summations necessary for formmg the normal equations are to be found in Table 6.

TABLE 6

X	Y	X^2	XY	Y^2	Y'	$Y - Y'$	$(Y - Y')^2$
0	1	0	0	1	1 646	−0 646	0 417316
1	3	1	3	9	2 022	0 978	0 956484
3	2	9	6	4	2 774	−0 774	0 599076
6	5	36	30	25	3 902	1 098	1 205604
8	4	64	32	16	4 654	−0 654	0 427716
18	15	110	71	55	14 998	0 002	3 606196

The normal equations are

$$5a + 18b = 15$$

$$18a + 110b = 71$$

and have the solutions

$$a = \tfrac{372}{226} = 1.646, \quad b = \tfrac{85}{226} = 0.376$$

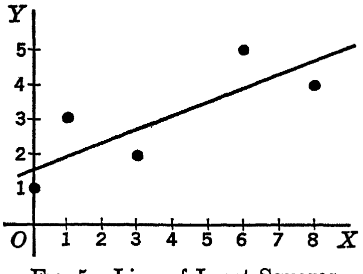


Fig 5.—Line of Least Squares.

The lme of least squares is therefore

$$Y' = 1.646 + 0\ 376X$$

Note that it can also be written in the form (8)

$$Y' - 3 = 0.376(X - 3.6)$$

The line is shown in Fig. 5. Using (10), we find that .

$$\Sigma(Y - Y')^2 = 55 - \tfrac{372}{226} \times 15 - \tfrac{85}{226} \times 71 = \tfrac{815}{226} = 3.6061947$$

This sum of squares has also been computed by the other method in Table 6, but it is readily seen that the use of formula (10) is much to be preferred.

19. Transformations. Certain types of equation can be reduced to linear form by a proper transformation of the variables. For example, some data, such as the numbers of bacteria in a colony at different times, conform to the *exponential* equation

$$Y' = AB^X \tag{13}$$

which may be written in the alternative form

$$Y' = Ae^{CX} \tag{14}$$

Taking logarithms of both sides of (13), we get

$$\log Y' = \log A + X \log B \tag{15}$$

or, if we set $\log Y' = y'$, $\log A = a$, $\log B = \mathrm{b}$,

$$y' = a + bX \tag{16}$$

We can now fit a line of least squares to the pairs of values $(X, y = \log Y)$ as in section 18. It should be noted that it is $\Sigma(y - y')^2$, not $\Sigma(Y - Y')^2$, which will be minimized. This sum of squares is given by

$$\Sigma(y - y')^2 = \Sigma y^2 - a\Sigma y - b\Sigma Xy \tag{17}$$

To discover whether a set of data conforms to the law (13) we may plot X and $\log Y$ on ordinary graph paper, or X and Y on semi-logarithmic paper, that is, paper on which one set of rulings is uniformly spaced and the other set logarithmically spaced. If in either case the points seem to lie along a straight line, we may fit a line of type (16).

In Table 7, column Y gives the number of bacteria per unit of

TABLE 7

X	Y	x	$y = \log 0\ 1Y$	xy
0	73	−2	0 8633	−1 7266
1	91	−1	0 9590	−0 9590
2	112	0	1 0492	0
3	131	1	1 1173	1 1173
4	162	2	1 2095	2 4190
Total .		0	5 1983	0 8507

volume existing in a culture at the end of X hours. Plotting X and Y on semi-logarithmic paper, as in Fig. 6, using the arithmetic or uniform scale for X and the logarithmic scale for Y, we

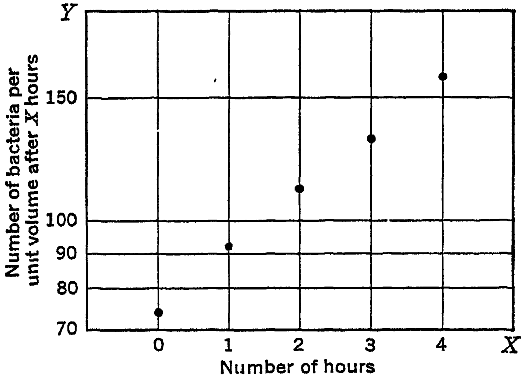


FIG. 6—Plot of X and Y of Table 7 on Semi-logarithmic Paper

see that the points lie approximately on a straight line. Figure 7 shows X and log Y plotted on ordinary graph paper.

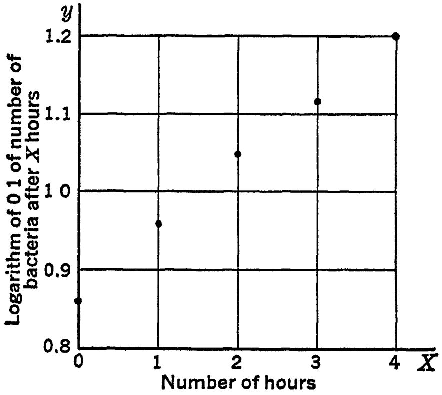


FIG. 7—Plot of X and y of Table 7.

The numerical work of fitting an exponential curve is shown in Table 7. In order to make the logarithms somewhat smaller we use log $0.1Y$ rather than log Y. This merely moves the decimal point in Y one place and reduces the characteristic of log Y by 1.

X = number of hours

Y = number of bacteria per unit volume after X hours

$$y' = \bar{y} + bx = \bar{y} + b(X - 2)$$

$$b = \frac{\Sigma x(y - \bar{y})}{\Sigma x^2} = \frac{\Sigma xy}{\Sigma x^2} = \frac{0.8507}{10} = 0.08507$$

$$y' = 1.03966 + 0\ 08507(X - 2)$$

$$= 0.86952 + 0.08507X$$

$$\log 0.1Y' = \log 7.40\bar{5} + X \log 1.216$$

$$0.1Y' = 7.40\bar{5}(1.216)^X$$

$$Y' = 74.0\bar{5}(1.216)^X$$

NOTE The minus sign over the 5 in the number $74.0\bar{5}$ indicates that this numbe1 is less than 74 05. This knowledge is useful if we wish to cut down the number of digits in the number Here the number to the nearest tenth is 74 0, not 74 1

The sum of squares of deviations is

$$\Sigma(y - y')^2 = \Sigma y^2 - 1.03966\Sigma y - 0.08507\Sigma xy$$

$$= 5.47704\,^* - 1.03966 \times 5\ 1983 - 0.08507 \times 0.8507$$

$$= 5.47704 - 5.40446 - 0.07237$$

$$= 0.0002$$

Another type of equation which can be reduced to linear form is

$$Y' = AX^B \tag{18}$$

By taking logarithms of both sides we reduce this to the form

$$\log Y' = \log A + B \log X \tag{19}$$

or, if we set $\log Y' = y'$, $\log A = a$, $\log X = x$,

$$y' = a + Bx \tag{20}$$

Again, the constants a and B can be determined by the method of least squares.

* Machine-calculated without recording the individual values of y^2.

To determine whether a set of data conforms to the law (18) we may plot log X and log Y on ordinary graph paper, or X and Y on logarithmic paper (both scales logarithmic), and note whether the points tend to lie on a straight line.

The sum of squares of deviations is

$$\Sigma(y - y')^2 = \Sigma y^2 - a\Sigma y - B\Sigma xy \tag{21}$$

20. Multiple regression. If we are concerned with more than two variables we may wish to estimate one of them from all the others. For example, we may wish to estimate Y from X_1 and X_2 by means of a *multiple regression equation*

$$Y' = b_0 + b_1X_1 + b_2X_2 \tag{22}$$

The coefficients b_1 and b_2 are the *partial regression coefficients* of Y on X_1 and X_2 respectively; b_1, for example, measures the average increase of Y per unit increase of X_1, when X_2 is held constant. Geometrically this equation represents a plane, and the ordinary procedure is to determine the b's so as to minimize the sum of squares of vertical deviations (assuming the Y-axis to be vertical) from the plane. We are led to the normal equations

$$\begin{aligned} b_0N + b_1\Sigma X_1 + b_2\Sigma X_2 &= \Sigma Y \\ b_0\Sigma X_1 + b_1\Sigma X_1{}^2 + b_2\Sigma X_1X_2 &= \Sigma X_1Y \\ b_0\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2{}^2 &= \Sigma X_2Y \end{aligned} \tag{23}$$

which have the solutions

$$\begin{aligned} b_0 &= \frac{1}{D}\begin{vmatrix} \Sigma Y & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1Y & \Sigma X_1{}^2 & \Sigma X_1X_2 \\ \Sigma X_2Y & \Sigma X_1X_2 & \Sigma X_2{}^2 \end{vmatrix}, \quad b_1 = \frac{1}{D}\begin{vmatrix} N & \Sigma Y & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1Y & \Sigma X_1X_2 \\ \Sigma X_2 & \Sigma X_2Y & \Sigma X_2{}^2 \end{vmatrix} \\ b_2 &= \frac{1}{D}\begin{vmatrix} N & \Sigma X_1 & \Sigma Y \\ \Sigma X_1 & \Sigma X_1{}^2 & \Sigma X_1Y \\ \Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2Y \end{vmatrix}, \quad D = \begin{vmatrix} N & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1{}^2 & \Sigma X_1X_2 \\ \Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2^2 \end{vmatrix} \end{aligned} \tag{24}$$

If we measure X_1, X_2, Y from their respective means, setting

$$x_1 = X_1 - \bar{X}_1, \quad x_2 = X_2 - \bar{X}_2, \quad y = Y - \bar{Y}$$

the first normal equation yields $b_0 = 0$ at once and the others reduce to

$$\begin{aligned} b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 &= \Sigma x_1y \\ b_1\Sigma x_1x_2 + b_2\Sigma x_2^2 &= \Sigma x_2y \end{aligned} \tag{25}$$

From these, or from (24), we find

$$\begin{aligned} b_1 &= \begin{vmatrix} \Sigma x_1y & \Sigma x_1x_2 \\ \Sigma x_2y & \Sigma x_2^2 \end{vmatrix} \div \begin{vmatrix} \Sigma x_1{}^2 & \Sigma x_1x_2 \\ \Sigma x_1x_2 & \Sigma x_3^2 \end{vmatrix} \\ b_2 &= \begin{vmatrix} \Sigma x_1^2 & \Sigma x_1y \\ \Sigma x_1x_2 & \Sigma x_2y \end{vmatrix} \div \begin{vmatrix} \Sigma x_1^2 & \Sigma x_1x_2 \\ \Sigma x_1x_2 & \Sigma x_2^2 \end{vmatrix} \end{aligned} \tag{26}$$

The terms such as $\Sigma x_1{}^2$, Σx_1x_2, and Σx_1y can be found from the relations

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}, \quad \Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}$$

For the sum of squares of deviations from the regression plane we have

$$\begin{aligned} \Sigma(Y - Y')^2 &= \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma X_1Y - b_2\Sigma X_2Y \\ &= \Sigma y^2 - b_1\Sigma x_1y - b_2\Sigma x_2y \end{aligned}$$

$$= \begin{vmatrix} \Sigma Y^2 & \Sigma Y & \Sigma X_1Y & \Sigma X_2Y \\ \Sigma Y & N & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1Y & \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1X_2 \\ \Sigma X_2Y & \Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2^2 \end{vmatrix} \div \begin{vmatrix} N & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1X_2 \\ \Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2^2 \end{vmatrix}$$

$$= \begin{vmatrix} \Sigma y^2 & \Sigma x_1y & \Sigma x_2y \\ \Sigma x_1y & \Sigma x_1^2 & \Sigma x_1x_2 \\ \Sigma x_2y & \Sigma x_1x_2 & \Sigma x_2^2 \end{vmatrix} \div \begin{vmatrix} \Sigma x_1^2 & \Sigma x_1x_2 \\ \Sigma x_1x_2 & \Sigma x_2^2 \end{vmatrix} \tag{27}$$

For the actual numerical solution of a set of normal equations it is better to use some such systematic method as the Doolittle method * or the method which we shall now illustrate by fitting a

* See Frederick C Mills, "Statistical Methods Applied to Economics and Business," Henry Holt & Co , New York, 1938, pp. 655–659.

linear regression of the form (22) to the data in the first three columns of Table 8.

TABLE 8

X_1	X_2	Y	X_1^2	X_1X_2	X_2^2	X_1Y	X_2Y
0	4	1	0	0	16	0	4
1	4	3	1	4	16	3	12
3	3	2	9	9	9	6	6
6	2	5	36	12	4	30	10
8	0	4	64	0	0	32	0
18	13	15	110	25	45	71	32

The normal equations and their solutions are shown below:

		Sum of coefficients
(A)	$5b_0 + 18b_1 + 13b_2 - 15 = 0$	21
(B)	$18b_0 + 110b_1 + 25b_2 - 71 = 0$	82
(C)	$13b_0 + 25b_1 + 45b_2 - 32 = 0$	51
(D) = 5(B) − 18(A)	$226b_1 - 109b_2 - 85 = 0$	32
(E) = 5(C) − 13(A)	$-109b_1 + 56b_2 + 35 = 0$	−18
(F) = 109(D) + 226(E)	$775b_2 - 1355 = 0$	−580
(G) = (F) − 775	$b_2 - 1\ 7484 = 0$	−0.7484
(D)	$226b_1 = 109 \times 1\ 7484 + 85 = 275.5756,$	$b_1 = 1\ 2194$
(E)	$-109b_1 = -56 \times 1\ 7484 - 35 = -132\ 9104,$	$b_1 = 1.2194$
(A)	$5b_0 = -18 \times 1.2194 - 13 \times 1\ 7484 + 15 = -29\ 6784,$	$b_0 = -5\ 9357$
(B)	$18b_0 = -110 \times 1\ 2194 - 25 \times 1\ 7484 + 71 = -106\ 8440,$	$b_0 = -5\ 9358$
(C)	$13b_0 = -25 \times 1\ 2194 - 45 \times 1\ 7484 + 32 = -77\ 1630,$	$b_0 = -5\ 9356$

The method of obtaining equations (D) and (E) is indicated in the solution. For example, (D) is obtained by multiplying (B) by 5, the coefficient of b_0 in (A), and (A) by −18, the negative of the coefficient of b_0 in (B), and adding the results, thus eliminating b_0. If it seems desirable, these extra equations, which may be symbolically written

$$(A') = -18(A) \quad \text{and} \quad (B') = 5(B)$$

may be inserted between (C) and (D). Similar equations may be inserted at the proper places in the solution. However, an experienced computer would probably prefer to obtain (D) and (E) directly, and this saves the labor of writing down the equations such as (A′) and (B′).

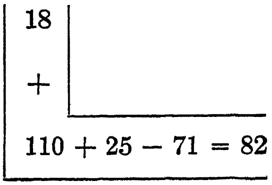
The column headed " Sum of coefficients " serves as an invaluable check on the correctness of the work at each stage of the solution. The sum of the coefficients in equation (A) is 21, that of the coefficients in (B) is 82. Since (D) is obtained by multiplying (B) by 5 and (A) by −18 and adding, the sum of the coefficients of (D) must be equal to 5 × 82 − 18 × 21 = 32. Similarly, the other steps may be checked by performing on the respective sums of coefficients the same operations that have been performed upon the equations.

It will be observed that in all the above equations the coefficients of the b's are symmetric about their principal diagonal. Because of this it is not necessary to rewrite those coefficients below this diagonal. A compact form of arranging the foregoing solution (as far as $b_2 = 1\ 7484$) is shown below.

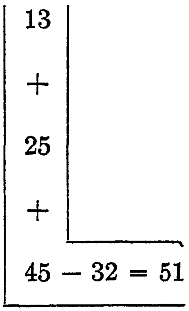
	b_0	b_1	b_2		s
(A)	5	18	13	−15	21
(B)		110	25	−71	82
(C)			45	−32	51
(D)		226	−109	−85	32
(E)			56	35	−18
(F)			775	−1355	−580
(G)			1	−1 7484	−0 7484

In calculating the " sum of coefficients " column, labeled s, one must realize that an L-shaped path must often be taken, since those coefficients below the diagonal have not been written down. This path has its corner, or turning point, at the number in the diagonal. Thus we have by way of illustration in the present example.

for equation (B)



for equation (C)



The regression equation is

$$Y' = -5.936 + 1.219X_1 + 1.748X_2$$

or

$$Y' - 3 = 1.219(X_1 - 3.6) + 1.748(X_2 - 2.6)$$

R. A. Fisher,* following Gauss, has given the following method of handling the normal equations: Let us replace the normal equations (23) by the three sets of equations

$$\begin{array}{rlll} & \text{For } j = 0, & 1, & 2 \\ c_{0j}N + c_{1j}\Sigma X_1 + c_{2j}\Sigma X_2 & = 1, & 0, & 0 \\ c_{0j}\Sigma X_1 + c_{1j}\Sigma X_1^2 + c_{2j}\Sigma X_1X_2 & = 0, & 1, & 0 \\ c_{0j}\Sigma X_2 + c_{1j}\Sigma X_1X_2 + c_{2j}\Sigma X_2^2 & = 0, & 0, & 1 \end{array} \qquad (28)$$

with $c_{ij} = c_{ji}$. (It is equivalent to replace equations (25) by

$$\begin{array}{ll} c_{11}\Sigma x_1^2 + c_{12}\Sigma x_1x_2 = 1 & c_{12}\Sigma x_1^2 + c_{22}\Sigma x_1x_2 = 0 \\ c_{11}\Sigma x_1x_2 + c_{12}\Sigma x_2^2 = 0 & c_{12}\Sigma x_1x_2 + c_{22}\Sigma x_2^2 = 1 \end{array} \qquad (29)$$

where the x's are, as usual, deviations from means.) If D is the determinant of the coefficients of (28), then $c_{ij} = D_{ij}/D$, where D_{ij} is the cofactor of the element of D which is the coefficient of c_{ij}, that is $(-1)^{i+j}$ times the minor obtained from D by striking out the row and the column occupied by the coefficient of c_{ij}.†

* "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh and London, section 29.

† See Maxime Bôcher, "Introduction to Higher Algebra," The Macmillan Co, New York, 1907.

In the present example we have the set-up and solution shown below.

	c_{0j}	c_{1j}	c_{2j}	$j = 0$	$j = 1$	$j = 2$	s_0	s_1	s_2
(A)	5	18	13	−1	0	0	35	36	36
(B)		110	25	0	−1	0	153	152	153
(C)			45	0	0	−1	83	83	82
(D)		226	−109	18	−5	0	135	112	117
(E)			56	13	0	−5	−40	−53	−58
(F)			775	4900	−545	−1130	5675	230	−355
(G)			1	6 3226	−0 7032	−1 4581	7 3226	2 9678	−0 4581

$$c_{02} = -\ 6\ 3226, \qquad c_{12} = 0\ 7032, \qquad c_{22} = 1.4581$$

(D) $226\, c_{01} = 109(-6\ 3226) - 18 = -\ 707\,1634,$ $\quad c_{01} = -\ 3.1290$

(D) $226\, c_{11} = 109 \times 0\ 7032 + 5 = 81\ 6488,$ $\quad c_{11} = 0\ 3613$

(E) $-\ 109\, c_{01} = -\ 56(-6\ 3226) - 13 = 341\ 0656,$ $\quad c_{01} = -\ 3.1290$

(E) $-\ 109\, c_{11} = -\ 56 \times 0\ 7032 + 0 = -\ 39\ 3792,$ $\quad c_{11} = 0\ 3613$

(A) $5\, c_{00} = -\ 18(-3\ 1290) - 13(-6\ 3226) + 1 = 139\ 5158,$ $\quad c_{00} = 27\ 9032$

(B) $18\, c_{00} = -110(-3\ 1290) - 25(-6\ 3226) + 0 = 502\ 2550,$ $\quad c_{00} = 27.9031$

Here we have solved all three sets, of three equations each, at the same time. The solution follows very closely that previously given. The constant term in each equation, however, is different, and naturally the s or " sum of coefficients " columns will be different. For example, (A) really consists of three equations, each corresponding to a fixed value of j. When $j = 0$, we have

$$5c_{00} + 18c_{01} + 13c_{02} - 1 = 0, \quad s = 35$$

When $j = 1$, we have

$$5c_{01} + 18c_{11} + 13c_{12} = 0, \quad s = 36$$

When $j = 2$, we have

$$5c_{02} + 18c_{12} + 13c_{22} = 0, \quad s = 36$$

The b's can be found from the c's by means of the relations

$$\begin{aligned} b_0 &= c_{00}\Sigma Y + c_{01}\Sigma X_1 Y + c_{02}\Sigma X_2 Y \\ b_1 &= c_{01}\Sigma Y + c_{11}\Sigma X_1 Y + c_{12}\Sigma X_2 Y \\ b_2 &= c_{02}\Sigma Y + c_{12}\Sigma X_1 Y + c_{22}\Sigma X_2 Y \end{aligned} \tag{30}$$

In the present problem we have

$$b_0 = 27.9032 \times 15 - 3\ 1290 \times 71 - 6.3226 \times 32 = -\ 5.9342$$

$$b_1 = -\ 3.1290 \times 15 + 0.3613 \times 71 + 0.7032 \times 32 = 1.2197$$

$$b_2 = -\ 6.3226 \times 15 + 0\ 7032 \times 71 + 1.4581 \times 32 = 1\ 7474$$

It will be noted that there are slight discrepancies between these values and those previously obtained. These are due to the fact that the c's have not been carried to a sufficient number of decimal places to make the values of the b's that we have just obtained quite as accurate as the previous values.

One advantage of the method just described is that, if we wish to find the regression equation for a new set of Y's with the same set of X_1's, X_2's, etc., we can make use of the c's already found and determine the b's from (30) with little extra labor. Suppose, for example, that we have 25 stations at which we have observed the first flowering dates of 10 species of plants over a period of 12 years. Let variables X_1 and X_2 be the altitude and latitude respectively of the stations, X_3 the year. If Y_1 is the first flowering date of the first species, Y_2 of the second species, and so on, we might want to determine ten regression equations (one for each species) of the type

$$Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Instead of deriving each of these separately we could derive the c_{ij} for the altitude and the latitude of the 25 stations and for the 12 years. These would be

$$\begin{array}{cccc} c_{00}, & c_{01}, & c_{02}, & c_{03} \\ & c_{11}, & c_{12}, & c_{13} \\ & & c_{22}, & c_{23} \\ & & & c_{33} \end{array}$$

since $c_{ij} = c_{ji}$, and from them we could find the 10 b's by using formulas similar to (30).

Another advantage of this method will be evident when we study the significance of regression coefficients.

It is not difficult to extend the foregoing discussion of multiple regression to the case of k independent variables $X_1, \ldots, X_k$. The equation of the regression may be written

$$Y' = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{31}$$

and the normal equations are

$$\begin{aligned}
b_0N + b_1\Sigma X_1 + b_2\Sigma X_2 + \cdots + b_k\Sigma X_k &= \Sigma Y \\
b_0\Sigma X_1 + b_1\Sigma X_1{}^2 + b_2\Sigma X_1X_2 + \cdots + b_k\Sigma X_1X_k &= \Sigma X_1Y \\
b_0\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2{}^2 + \cdots + b_k\Sigma X_2X_k &= \Sigma X_2Y \\
\cdot \quad \cdot \qquad\qquad\qquad\qquad\qquad \cdot \quad \cdot & \\
b_0\Sigma X_k + b_1\Sigma X_1X_k + b_2\Sigma X_2X_k + \cdots + b_k\Sigma X_k{}^2 &= \Sigma X_kY
\end{aligned} \tag{32}$$

The sum of squares of deviations is

$$\Sigma(Y - Y')^2 = \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma X_1Y - b_2\Sigma X_2Y - \cdots - b_k\Sigma X_kY \tag{33}$$

21. Curvilinear regression. The X_i do not actually have to be independent. For example, if we wish to fit a regression equation of the type

$$Y' = b_0 + b_1X + b_2Z + b_3X^2 + b_4XZ + b_5Z^2$$

we can set $X = X_1$, $Z = X_2$, $X^2 = X_3$, $XZ = X_4$, $Z^2 = X_5$, and proceed as above.

One important case is that in which the regression function is a polynomial

$$Y' = b_0 + b_1X + b_2X^2 + \cdots + b_kX^k \tag{34}$$

For purposes of illustration we shall fit a second-degree equation (geometrically a parabola) to the points of section 18 (see Table 6). The normal equations are

$$\begin{aligned}
b_0N + b_1\Sigma X + b_2\Sigma X^2 &= \Sigma Y \\
b_0\Sigma X + b_1\Sigma X^2 + b_2\Sigma X^3 &= \Sigma XY \\
b_0\Sigma X^2 + b_1\Sigma X^3 + b_2\Sigma X^4 &= \Sigma X^2Y
\end{aligned} \tag{35}$$

The various summations needed will be found in Table 9.

TABLE 9

X	X^2	X^3	X^4	Y	XY	X^2Y	Y^2
0	0	0	0	1	0	0	1
1	1	1	1	3	3	3	9
3	9	27	81	2	6	18	4
6	36	216	1296	5	30	180	25
8	64	512	4096	4	32	256	16
18	110	756	5474	15	71	457	55

The solution is shown below.

	b_0	b_1	b_2		s	
(A)	5	18	110	− 15	118	
(B)		110	756	− 71	813	
(C)			5,474	−457	5,883	
(D)		226	1,800	− 85	1,941	5 (B) − 18 (A)
(E)		1800	15,270	−635	16,435	5 (C) − 110 (A)
(F)			211,020	9490	220,510	226 (E) − 1800 (D)
(G)			1	0 04497	1 04497	(F) ÷ 211,020

(G) $b_2 = -\ 0.04497$

(D) $226\, b_1 = -\ 1{,}800(-0.04497) + 85 = 165.9460,$ $\quad b_1 = 0\ 73427$

(E) $1800\, b_1 = -\ 15{,}270(-0\ 04497) + 635 = 1321\ 6919,$ $\quad b_1 = 0\ 73427$

(A) $5\, b_0 = -\ 18 \times 0.73427 - 110(-0\ 04497) + 15 = 6\ 72984,$ $\quad b_0 = 1.345968$

(B) $18\, b_0 = -110 \times 0.73427 - 756(-0\ 04497) + 71 = 24\ 22762,$ $\quad b_0 = 1\ 345979$

The regression equation is therefore

$$Y' = 1.3460 + 0.73427X - 0.04497X^2 \tag{36}$$

The sum of squares of vertical deviations from the regression curve is

$$\begin{aligned}\Sigma(Y - Y')^2 &= \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma XY - b_2\Sigma X^2Y \\ &= 55 - 1.3460 \times 15 - 0\ 73427 \times 71 + \\ &\qquad 0.04497 \times 457 = 3\ 22812\end{aligned}$$

Comparing this value with that found for the sum of squares of deviations from the straight line fitted in section 18, viz., 3.60619, we see that although it is less, as is to be expected, it is very little less. That is, the parabola fits the points very little better than the straight line. The small value of b_2 means that the X^2 term has little influence on our estimated value of Y.

When the degree of the polynomial to be fitted has not been decided upon in advance, it is possible to fit successively a constant (the mean), a linear function, a quadratic, and so on, each function being obtained from the preceding by the addition of a new term. The technique of this process is fully explained by Fisher.*

EXERCISES

1. The length of a spiral spring under various loads is given in the following table·

Load in grams, X.	5	10	15	20	25	30
Length in centimeters, Y	7 25	8 12	8 95	9 90	10 9	11 8

(*a*) Plot these values, and find the equation of a least squares line of the form $Y' = a + bX$ (*b*) Find $\Sigma(Y - Y')^2$

2. Table J gives the intelligence quotients and the scores on a reading

TABLE J

INTELLIGENCE QUOTIENTS AND SCORES IN A READING VOCABULARY TEST OF A GROUP OF FIFTH-GRADE PUPILS

Pupil	I. Q	Reading vocabulary	Pupil	I Q	Reading vocabulary
1	140	64	14	112	51
2	139	54	15	105	44
3	135	55	16	105	35
4	134	56	17	103	44
5	126	46	18	103	29
6	126	51	19	96	33
7	124	61	20	94	30
8	118	42	21	93	31
9	115	41	22	92	33
10	114	47	23	76	14
11	114	43	24	70	12
12	113	56	25	66	15
13	113	48			

* "Statistical Methods for Research Workers," sections 27–28.1.

vocabulary test of a group of fifth-grade pupils (Yates, A M. Thesis, Washington University, 1937) (*a*) Make a scatter diagram for these data (*b*) Find the equation of the line of regression of reading vocabulary on I Q (*c*) Find the sum of squares of deviations from this line

3. Plot the data of Table K, and fit a straight line of trend to them.

TABLE K

Rate per 100,000 Population of Automobile Fatalities in the United States (*Statistical Abstract of the United States*)

Year	Rate	Year	Rate
1911	2 2	1923	14 7
1912	2 9	1924	15 5
1913	3 9	1925	17 1
1914	4 3	1926	18 0
1915	5 9	1927	19 6
1916	7 3	1928	20 8
1917	9 0	1929	23 3
1918	9 3	1930	24 5
1919	9 4	1931	25 2
1920	10 4	1932	21 9
1921	11 4	1933	23 3
1922	12 4	1934	26 8

TABLE L

Population of the United States

Year	Population in millions	Year	Population in millions
1790	3 9	1870	38 6
1800	5 3	1880	50 2
1810	7 2	1890	62 9
1820	9 6	1900	76 0
1830	12 9	1910	92 0
1840	17 1	1920	105 7
1850	23 2	1930	122 8
1860	31 4		

4. (*a*) Plot the population of the United States for the census years (Table L) on ordinary graph paper. (*b*) Plot these same data on semi-logarithmic paper, or plot the logarithms of the population against equally spaced time values (*c*) Fit a straight line of trend to the period 1880–1930, and calculate the theoretical values of the population for the census years of this period (*d*) Fit an exponential curve to the period 1790–1900, and calculate the theoretical values

5. (*a*) Fit a parabola to the cotton production data of Table M. (*b*) Draw the parabola and plot the data (*c*) Find the sum of squares of deviations from the curve

TABLE M

COTTON PRODUCTION IN THE UNITED STATES

(*Survey of Current Business*)

Year	Production (millions of bales)	Year	Production (millions of bales)
1931	17 1	1935	10 6
1932	13 0	1936	12 4
1933	13 0	1937	18 2
1934	9 6		

6. The following pairs of values of the pressure p and volume v of a gas were measured in a laboratory experiment Find an equation of the type $p' = Av^B$ connecting the two variables.

v	1 5	1 7	2	2 6	3	3 6	4 2	5
p	106	89 0	70 8	49 0	40 0	30 0	24 5	19 5

7. Fit an exponential curve to the data of Table 2, page 3.

8. (*a*) Find a multiple regression equation of grade-point average on the other variables of Table N (*b*) Find the sum of squares of deviations from the regression function.

TABLE N

Scores of 100 Students in an Intelligence Test, Reading Comprehension, and Reading Rate, and Grade-Point Average for First Semester in University

Intelligence	Reading comprehension	Reading rate	Grade-point average	Intelligence	Reading comprehension	Reading rate	Grade-point average
275	153	29	1 2	295	186	41	2 4
181	132	22	0 0	152	107	18	0 6
152	144	34	0 6	214	198	45	0 2
162	134	31	1 0	171	139	29	0 0
158	145	34	1 2	131	111	28	1 0
228	179	47	2 8	178	149	38	0 6
273	172	38	2 2	225	143	25	1 0
246	148	36	0 8	141	122	26	0 4
235	166	29	2 0	116	83	22	0 0
131	124	26	0 6	173	144	37	2 6
247	158	25	1 0	230	179	37	2 6
187	90	22	0 6	174	114	24	1 8
150	120	29	1 2	177	133	32	0 0
234	183	41	1 2	210	151	26	0 4
215	178	41	0 0	236	132	29	1 8
237	139	24	1 4	198	160	34	0 8
182	159	31	0 6	217	178	38	1 0
177	152	29	0 0	143	153	40	0 2
165	161	48	1 0	186	144	27	2 8
317	192	30	2 0	233	172	44	1 4
202	125	28	2 0	136	109	32	0 2
161	131	17	0 8	183	142	26	0 4
181	130	27	0 6	223	167	50	1 4
176	126	26	0 4	106	116	24	0 0
271	178	25	3 0	211	128	18	0 8
174	148	22	0 6	151	93	20	0 4
161	136	29	0 2	231	171	26	2 2
231	156	32	1 4	135	152	26	1 4
229	174	33	2 0	146	154	19	1 2
152	116	23	0.7	227	149	35	1 4
97	113	20	1 0	204	160	26	1 4
182	165	38	1 4	223	109	18	1 4
247	168	27	2 2	142	101	22	0 8
232	191	44	0 4	176	127	22	0 8
246	155	29	1 2	238	164	27	2 6
180	101	25	1 0	268	177	40	2 6
247	158	25	1 0	163	139	33	0 2
188	148	18	0 0	195	140	38	0 0
233	144	27	1 2	184	143	32	0 8
154	113	29	1 2	192	119	22	0 8
270	194	50	2 0	121	100	34	0 6
220	175	24	1 2	316	176	42	2 6
206	136	33	0 8	234	183	41	1 2
170	185	40	0 6	146	112	18	0 6
240	162	50	1 6	261	175	35	2 6
192	150	32	0 6	175	158	30	1 2
157	119	30	0 4	233	182	34	1 6
223	154	29	0 4	261	156	25	2 4
221	152	14	1 2	242	187	49	1 4
262	162	44	0 6	134	170	48	0 8

CHAPTER IV

CORRELATION

22. Coefficient of correlation. As we realize from the discussion of regression, there are many instances in which an increase in one variable is in general accompanied by an increase in another. In other words, large values of one variable tend to be associated with large values of a second, while small values of the first tend to be associated with small values of the second, without the existence of a strict mathematical relationship between the two. In such a case the variables are said to be *correlated.* Of course an increase in one variable may be accompanied by a decrease in the other, that is, large values of each variable may tend to be associated with small values of the other. This situation is called *inverse* correlation.

The most widely used measure of correlation is the *coefficient of correlation.* For the case of N pairs of values of the variables X and Y the coefficient of correlation may be defined as

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{[\Sigma(X - \bar{X})^2 \cdot \Sigma(Y - \bar{Y})^2]^{\frac{1}{2}}} = \frac{\Sigma xy}{(\Sigma x^2 \cdot \Sigma y^2)^{\frac{1}{2}}} = \frac{\Sigma xy}{N\sigma_X\sigma_Y} \qquad (1)$$

in which σ_X and σ_Y are the standard deviations of X and Y respectively. It is to be noted that it is entirely symmetric with respect to X and Y. Moreover, it is entirely independent of the units in which X and Y are expressed, and may have any value between -1 and 1. If large values of the two variables tend to be associated, the factors $X - \bar{X}$ and $Y - \bar{Y}$ in the numerator of r will usually have like signs and the sum of products will accumulate, giving a value of r near to 1. If there is little correlation between X and Y the products $(X - \bar{X})(Y - \bar{Y})$ will sometimes have a positive sign, sometimes a negative, thus tending to neutralize each other and resulting in a value of r near zero. Inverse correlation yields a negative value of r. The coefficient r will have the value 1 or -1 only in the case of perfect correlation, that is, when there exists a definite linear relation between X and Y.

For the actual computation of r, (1) may be placed in the form

$$r = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)/N}{[\Sigma X^2 - (\Sigma X)^2/N]^{1/2}[\Sigma Y^2 - (\Sigma Y)^2/N]^{1/2}} \tag{2}$$

in which the variables may be measured from any origin and in terms of any unit. If one of the variables is measured from its mean, r assumes one of the following simpler forms:

$$r = \frac{\Sigma xY}{(\Sigma x^2)^{1/2}[\Sigma Y^2 - (\Sigma Y)^2/N]^{1/2}} = \frac{\Sigma Xy}{[\Sigma X^2 - (\Sigma X)^2/N]^{1/2}(\Sigma y^2)^{1/2}} \tag{3}$$

If both variables are measured from their means, the resulting form of r is shown in (1)

Let us compute r for the following pairs of numbers:

X	0, 1, 3, 6, 8
Y	1, 3, 2, 5, 4

From Table 6 we get

$$\Sigma X = 18, \quad \Sigma Y = 15, \quad \Sigma X^2 = 110, \quad \Sigma XY = 71, \quad \Sigma Y^2 = 55$$

Hence, using (2), we find

$$r = \frac{71 - 18 \times 15/5}{[110 - (18)^2/5]^{1/2}[55 - (15)^2/5]^{1/2}}$$

$$= \frac{17}{(45.2 \times 10)^{1/2}} = 0.80$$

23. Connection between correlation and regression. The equation of the line of regression of Y on X (see section 18) is

$$y' = b_{YX}\,x, \quad b_{YX} = \frac{\Sigma xy}{\Sigma x^2} \tag{4}$$

Similarly, the equation of the line of regression of X on Y is

$$x' = b_{XY}\,y, \quad b_{XY} = \frac{\Sigma xy}{\Sigma y^2} \tag{5}$$

From (4), (5), and (1) we see that

$$b_{XY}\,b_{YX} = r^2 \tag{6}$$

and that r is consequently equal to the square root of the product of the two regression coefficients, that is, to their geometric mean. These regression coefficients will both have the same sign, and r should be given their common sign.

It is also easily shown that

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}, \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} \tag{7}$$

Another instructive way of regarding the correlation coefficient is as follows: Suppose we have fitted a line of regression, $Y' = a + bX$, to a set of points. The points will cluster more closely about this line than they will about the horizontal line drawn through the mean of the Y's. That is, unless $b = 0$, the sum of squares of deviations of the Y's from the estimates Y' will be less than the sum of squares of deviations from the mean. The quotient

$$\frac{\Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2} \tag{8}$$

will therefore be less than 1. The more closely the points are clustered about the regression line the smaller will the numerator be. The quotient (8) is small if X and Y are closely correlated, and approaches its maximum value 1 if they are practically independent. Its value, which varies between 0 and 1, is then a sort of inverse measure of the correlation. If it were subtracted from 1 it should give a measure of the correlation. Actually it can be shown that

$$r^2 = 1 - \frac{\Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2} \tag{9}$$

$$= \frac{a\Sigma Y + b\Sigma XY - (\Sigma Y)^2/N}{\Sigma Y^2 - (\Sigma Y)^2/N} \tag{10}$$

The sign to be prefixed to r is that of the regression coefficient b.

Formula (9) is most readily proved if we use deviations from the mean. Thus, for the right side of the equation we can write

$$1 - \frac{\Sigma(y - y')^2}{\Sigma y^2} = 1 - \frac{\Sigma y^2 - b\Sigma xy}{\Sigma y^2}$$

$$= 1 - 1 + b\frac{\Sigma xy}{\Sigma y^2} = \frac{(\Sigma xy)^2}{\Sigma x^2 \cdot \Sigma y^2} = r^2$$

Formula (10) follows from (9) above and from (10) of section 18 (page 29).

Formula (9) may, of course, be used to find the sum of squares of deviations from the line of regression if r is known.

To the data of section 22 we have previously (section 18) fitted a regression line $Y' = a + bX$, finding $a = 1.646$, $b = 0.376$. Also we have, from section 22 or from Table 6,

$$\Sigma Y = 15, \quad \Sigma XY = 71, \quad \Sigma Y^2 = 55$$

Thus from (10)

$$r^2 = \frac{1.646 \times 15 + 0\ 376 \times 71 - (15)^2/5}{55 - (15)^2/5}$$

$$= \frac{24.690 + 26.696 - 45}{10} = 0.6386$$

$$r = 0.80$$

24. Correlation table. When the pairs of values of two variables are numerous they are often grouped into a *correlation table,* which is a table of double classification. Table 10 is a correlation table, the data, however, being fictitious to admit of an easy illustration of how r may be calculated from such a table. The numbers in the body of the table are frequencies. For example, the number 5 near the middle of the table indicates that there are five individuals whose heights are between 66 and 69 inches and whose weights are between 125 and 150 pounds.

TABLE 10

CORRELATION TABLE OF HEIGHTS AND WEIGHTS OF A GROUP OF MEN

Weight in pounds	Height in inches				
	60–63	63–66	66–69	69–72	72–75
100–125	2	1			
125–150	2	3	5	1	
150–175	. . .	2	4	1	2
175–200	.		1	1	

Since the value of r is quite independent of the units in which either variable is measured, we find it most convenient to measure each in terms of its class interval, and from the middle of some centrally located class as origin, as shown in Table 11.

TABLE 11

COMPUTATION OF r

Y \ X	−2	−1	0	1	2	f_Y	Yf_Y	Y^2f_Y	$(Y+1)^2f_Y$
−1	2 2 −1	1 1 0				3	−3	3	0
0	2 0 −2	3 0 −1	5 0 0	1 0 1		11	0	0	11
1		2 −1 −2	4 0 −1	1 1 0	2 2 1	9	9	9	36
2		.	1 0 −2	1 2 −1		2	4	8	18
f_X	4	6	10	3	2	25	10	20	65
Xf_X	−8	−6	0	3	4	−7			
X^2f_X	16	6	0	3	8	33			
$(X+1)^2f_X$	4	0	10	12	18	44			

TABLE 11A

XY	f_{XY}	XYf_{XY}
−1	2	−2
0	16	0
1	2	2
2	5	10
	25	10

TABLE 11B

$(X-Y)^2$	f_{XY}	$(X-Y)^2f_{XY}$
0	7	0
1	13	13
4	5	20
	25	33

Charlier Checks

$$\begin{array}{rlcrlcrl}
\Sigma \mathrm{X}^2 & = 33 & \qquad & \Sigma Y^2 & = 20 & \qquad & \Sigma X^2 & = 33 \\
2\Sigma X & = -14 & & 2\Sigma Y & = 20 & & -2\Sigma XY & = -20 \\
N & = \underline{25} & & N & = \underline{25} & & \Sigma Y^2 & = \underline{20} \\
\Sigma(X+1)^2 & = 44 & & \Sigma(Y+1)^2 & = 65 & & \Sigma(X-Y)^2 & = 33
\end{array}$$

$$\begin{aligned}
r &= \frac{\Sigma XY - \Sigma X \cdot \Sigma Y/N}{[\Sigma X^2 - (\Sigma X)^2/N]^{\frac{1}{2}}[\Sigma Y^2 - (\Sigma Y)^2/N]^{\frac{1}{2}}} \\
&= \frac{10 - (-7)\ 10/25}{(33 - 49/25)^{\frac{1}{2}}(20 - 100/25)^{\frac{1}{2}}} \\
&= \frac{12\ 80}{(31\ 04 \times 16)^{\frac{1}{2}}} = 0\ 57
\end{aligned}$$

In this table the frequencies are shown in the middle of the various compartments. The small number in the lower left-hand corner of a compartment is the product of X and Y for that compartment; the number in the lower right-hand corner is the difference $X - Y$. The first of these is used in computing the sum of products, ΣXY; the second, in making a third Charlier check, viz.,

$$\Sigma(X - Y)^2 = \Sigma X^2 - 2\Sigma XY + \Sigma Y^2 \tag{11}$$

Table 11A shows the computation of ΣXY; Table 11B shows the computation of $\Sigma(X - Y)^2$. In the latter the frequencies corresponding to $(X - Y)^2 = (\pm 1)^2$ can, of course, be grouped together, as can those corresponding to $(X - Y)^2 = (\pm 2)^2$.

To use formula (10) in the computation of r we must fit a line of least squares, $Y' = a + bX$. The normal equations with their solutions are

$$\begin{aligned}
25a - 7b &= 10 \qquad & a &= 0.51546 \\
-7a + 33b &= 10 \qquad & b &= 0.41237
\end{aligned}$$

and from (10) we have

$$\begin{aligned}
r^2 &= \frac{0.51546 \times 10 + 0.41237 \times 10 - 4}{20 - 4} = 0.3299 \\
r &= 0.57
\end{aligned}$$

25. Correlation ratio. We have seen in section 23 that the square of the coefficient of correlation can be obtained by subtracting from unity the fraction whose numerator is the sum of squares of deviations of the Y's from the line of regression and whose denominator is the sum of squares of deviations of the Y's from their mean. If the regression of Y on X is not linear, that is, if the straight line of regression does not pass near to the points of the scatter diagram, we can get a better measure of the correlation by means of the *correlation ratio.*

Suppose that instead of fitting a straight line of regression to the data of a correlation table we find the mean of each column, plot these means on a graph, and then draw a broken line through the resulting points. See Fig. 8, which illustrates the data of the correlation table discussed in section 24. The dots in this figure are the individuals; the small circles indicate the means of the respective columns. By analogy with the square of the correlation coefficient, we define the square of the correlation ratio to be 1 minus the fraction whose numerator is the sum of squares of deviations from the means of the columns and whose denominator is the sum of squares of deviations from the general mean.

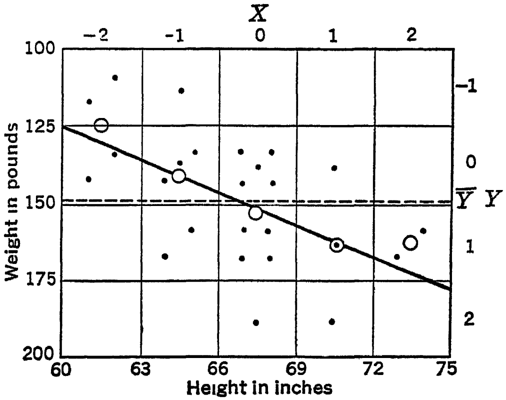


FIG 8—Correlation Diagram (Note that the Y-axis is positive downward and that consequently the regression line slopes downward on the graph although the coefficient of regression is positive)

To express the correlation ratio analytically, suppose that we have a correlation table consisting of k columns. Suppose that in the column corresponding to a fixed X there are N_X values of Y,

$$Y_{X1}, Y_{X2}, \dots, Y_{Xi}, \dots, Y_{XN_X}$$

whose mean value is $\bar{Y}_X$ If $\bar{Y}$ is the general mean of the Y's, the correlation ratio of Y on X is the square root of

$$\eta_{YX}^2 = 1 - \frac{\displaystyle\sum_{X=X_1}^{X_k} \sum_{i=1}^{N_X} (Y_{X_i} - \bar{Y}_X)^2}{\displaystyle\sum_{X=X_1}^{X_k} \sum_{i=1}^{N_X} (Y_{X_i} - \bar{Y})^2} \tag{12}$$

This can be shown to be equal to the ratio of the weighted sum of squares of deviations of the means of columns from the general mean (the weights being the numbers in the columns) to the sum of squares of deviations from the general mean, that is,

$$\eta_{YX}^2 = \frac{\Sigma N_X(\bar{Y}_X - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} \tag{13}$$

For purposes of computation (13) is superior to (12), although the following equivalent form is perhaps still better:

$$\eta_{YX}^2 = \left[\sum_{X=X_1}^{X_k} \frac{\left(\displaystyle\sum_{i=1}^{N_X} Y_{X_i}\right)^2}{N_X} - \frac{(\Sigma Y)^2}{N}\right] \div \left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right] \tag{14}$$

Here $N = N_1 + \ldots + N_k$ = total frequency, and ΣY and ΣY^2 indicate summation over the entire table, viz., for both the index i and the index X. That is, the denominators in (14), (12), and (13) are the same.

We shall illustrate below the calculation of η from the correlation table used above (Table 10). It seems more appropriate here to use subscripts $-2, -1, 0, 1, 2$ rather than $1, 2, 3, 4, 5$.

1st column ($X = -2$)

Y	f	Yf	Y^2f
-1	2	-2	2
0	2	0	0
	4	-2	2

$$N_{-2} = 4$$

$$\bar{Y}_{-2} = -\,2/4 = -\,0.5$$

$$\Sigma(Y - \bar{Y}_{-2})^2 = 2 - (-2)^2/4 = 1$$

2nd column ($X = -1$)

Y	f	Yf	Y^2f
−1	1	−1	1
0	3	0	0
1	2	2	2
	6	1	3

$$N_{-1} = 6$$

$$\overline{Y}_{-1} = 1/6 = 0.1\dot{6}$$

$$\Sigma(Y - \overline{Y}_{-1})^2 = 3 - 1^2/6 = 2.8\dot{3}$$

3rd column ($X = 0$)

Y	f	Yf	Y^2f
0	5	0	0
1	4	4	4
2	1	2	4
	10	6	8

$$N_0 = 10$$

$$\overline{Y}_0 = 6/10 = 0.6$$

$$\Sigma(Y - \overline{Y}_0)^2 = 8 - 6^2/10 = 4.4$$

4th column ($X = 1$)

Y	f	Yf	Y^2f
0	1	0	0
1	1	1	1
2	1	2	4
	3	3	5

$$N_1 = 3$$

$$\overline{Y}_1 = 3/3 = 1$$

$$\Sigma(Y - \overline{Y}_1)^2 = 5 - 3^2/3 = 2$$

5th column ($X = 2$)

Y	f	Yf	Y^2f
1	2	2	2

$$N_2 = 2$$

$$\overline{Y}_2 = 1$$

$$\Sigma(Y - \overline{Y}_2)^2 = 0$$

$$\sum_{X=X_1}^{X_k} \frac{\left(\sum_{i=1}^{N_Y} Y_{X_i}\right)^2}{N_X} - \frac{(\Sigma Y)^2}{N}$$

$$= \frac{(-2)^2}{4} + \frac{1^2}{6} + \frac{6^2}{10} + \frac{3^2}{3} + \frac{2^2}{2} - \frac{(-2+1+6+3+2)^2}{25}$$

$$= 1 + 0\,1\dot{6} + 3.6 + 3 + 2 - 4 = 5.7\dot{6}$$

$$\Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 2 + 3 + 8 + 5 + 2 - \frac{(-2+1+6+3+2)^2}{25}$$

$$= 20 - 4 = 16$$

From (14), $\eta^2_{YX} = \dfrac{5\,7\dot{6}}{16} = 0.3604,\ \eta_{YX} = 0.60$

With a calculating machine, ΣY^2 can be computed directly, in which case the columns headed Y^2f are unnecessary. These columns are, however, necessary if we use formula (12). Ordinarily the correlation ratio would not be calculated by (12), but to emphasize the analogy between the correlation ratio and the correlation coefficient, and also to verify the equivalence of (12) and (13) or (14), we shall carry through the calculation for this particular example. We find

$$\sum_{X=X_1}^{X_k} \sum_{i=1}^{N_X} (Y_{X_i} - \bar{Y}_X)^2 = 1 + 2.83 + 4.4 + 2 + 0 = 10.23$$

$$\sum_{X=X_1}^{X_k} \sum_{i=1}^{N_X} (Y_{X_i} - \bar{Y})^2 = 16, \text{ as before.}$$

From (12), $\eta^2_{YX} = 1 - \dfrac{10.23}{16} = 1 - 0.6396 = 0.3604$

$$\eta_{YX} = 0.60$$

If we had used the means of rows instead of the means of columns we should have found η_{XY}, the correlation ratio of X on Y. The formula for η_{XY} is obtained from any of the preceding formulas for η_{YX} by interchanging X and Y. It should be noted that in general η_{XY} is different from η_{YX}, although $r_{XY} = r_{YX} = r$. We shall always consider η as positive. It is necessarily larger than (in exceptional cases equal to) the numerical value of r, for the sum of squares of deviations from the means of arrays* will be less than the sum of squares of deviations from the line of regression (unless the means of arrays lie exactly on a straight line), and consequently there is a smaller fraction to be subtracted from 1 in the case of the correlation ratio than there is in the case of the correlation coefficient, leaving a larger remainder. In the preceding example we found $r = 0\ 57$, $\eta = 0\ 60$.

26. Relation between correlation coefficient and correlation ratio. Fréchet † has pointed out that the correlation coefficient computed from a correlation table can be separated into the product of two factors, thus:

$$r = \frac{\Sigma N_X(X - \bar{X})(\bar{Y}_X - \bar{Y})}{[\Sigma N_X(X - \bar{X})^2]^{\frac{1}{2}}[\Sigma N_X(\bar{Y}_X - \bar{Y})^2]^{\frac{1}{2}}} \cdot \frac{[\Sigma N_X(\bar{Y}_X - \bar{Y})^2]^{\frac{1}{2}}}{[\Sigma(Y - \bar{Y})^2]^{\frac{1}{2}}} \quad (15)$$

The first factor is the coefficient of correlation obtained by associating with each value of X the mean of the corresponding values of Y, weighted with the number of such Y's; the second is the correlation ratio η_{YX}. The first factor depends only on the line joining the means of columns, approaching 1 if this line approximates a straight line. It is not affected by the dispersion of the points in a given column. The correlation ratio, on the contrary, is not affected when the partial distributions of the columns are displaced by shifting the means, that is, by deforming the curve (or broken line) of the means. Thus the first factor does not depend on the closeness of the relation between X and Y. On the other hand, the value of the correlation ratio may be near 1 either because the points corresponding to a given X are near the mean of that column or because there is a small number of Y's for each value of X. Both these factors have to be near 1 to yield a value of r near 1.

We shall illustrate the separation of r into the two factors by using the data of the correlation table given above. To facilitate

* *Array* is a generic term for *row* or *column*

† Maurice Fréchet, "Sur le coefficient, dit de corrélation et sur la corrélation en général," *Revue de l'Institut International de Statistique*, vol 4, 1933, pp 1–8.

the computation we shall first throw the numerator of the first factor into a different form. It can easily be shown that

$$\Sigma N_X(X - \bar{X})(\bar{Y}_X - \bar{Y}) = \Sigma X N_X \bar{Y}_X - \frac{\Sigma X \cdot \Sigma Y}{N} \tag{16}$$

Using the values obtained when calculating η, we find that the foregoing is equal to

$$-2(-2)-1\cdot 1 + 0\cdot 6 + 1\cdot 3 + 2\cdot 2 - \frac{(-7)10}{25} = 12.8$$

In the calculations of r and η respectively we have already found

$$\Sigma N_X(X - \bar{X})^2 = 31.04, \quad \Sigma N_X(\bar{Y}_X - \bar{Y})^2 = 5.7\dot{6}$$

Consequently

$$r = \frac{12\ 8}{(31.04 \times 5.7\dot{6})^{1/2}} \eta_{YX} = 0.949 \times 0.60 = 0.57$$

which is the value previously found.

27. Index of correlation. When a curve $Y' = f(X)$ is fitted to a set of data we may define the *index of correlation* of Y on X as the square root of

$$R^2_{YX} = 1 - \frac{\Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2} \tag{17}$$

The index of correlation will be numerically greater than or equal to the coefficient of correlation. It can be used in connection with ungrouped data or grouped data The correlation ratio can be used only in connection with data which have been grouped into a correlation table, or which have more than one value of Y corresponding to a given value of X.

In section 21 we fitted a second-degree polynomial, obtaining the curvilinear regression equation

$$Y' = 1.3460 + 0.73427X - 0.04497X^2$$

In that section we also found

$$\Sigma Y = 15, \quad \Sigma Y^2 = 55, \quad \Sigma XY = 71, \quad \Sigma X^2 Y = 457$$

Thus

$$
\begin{aligned}
R^2_{YX} &= 1 - \frac{\Sigma(Y - Y')^2}{\Sigma(Y - \overline{Y})^2} \\
&= 1 - \frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY - c\Sigma X^2Y}{\Sigma Y^2 - (\Sigma Y)^2/N} \\
&= \frac{a\Sigma Y + b\Sigma XY + c\Sigma X^2Y - (\Sigma Y)^2/N}{\Sigma Y^2 - (\Sigma Y)^2/N} \\
&= \frac{1.3460 \times 15 + 0.73427 \times 71 - 0.04497 \times 457 - (15)^2/5}{55 - (15)^2/5} \\
&= \frac{6.77188}{10} = 0.677188 \\
R_{YX} &= 0.823
\end{aligned}
$$

The coefficient of correlation was previously found to be 0.80, and we see that the index of correlation is slightly larger.

28. Multiple correlation. If we have found a linear regression equation of X_1 on $X_2, \ldots, X_k$,

$$X_1' = b_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

then by analogy to the simple coefficient of correlation (see equation (9), section 23, page 49) we define the *coefficient of multiple correlation* between X_1 and $X_2, \ldots, X_k$ to be the square root of

$$r^2_{1\,23\cdot\ k} = 1 - \frac{\Sigma(X_1 - X_1')^2}{\Sigma(X_1 - \overline{X}_1)^2} \tag{18}$$

It will be remembered (cf. (33), section 20, page 41) that

$$\Sigma(X_1 - X_1')^2 = \Sigma X_1^2 - a_1\Sigma X_1 - b_{12}\Sigma X_1X_2 - \cdots - b_{1k}\Sigma X_1X_k \tag{19}$$

$$
= \begin{vmatrix}
N & \Sigma X_1 & \Sigma X_2 & \cdots & \Sigma X_k \\
\Sigma X_1 & \Sigma X_1^2 & \Sigma X_1X_2 & \cdots & \Sigma X_1X_k \\
\Sigma X_2 & \Sigma X_1X_2 & \Sigma X_2^2 & \cdots & \Sigma X_2X_k \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\Sigma X_k & \Sigma X_1X_k & \Sigma X_2X_k & \cdots & \Sigma X_k^2
\end{vmatrix}
$$

$$
\div \begin{vmatrix}
N & \Sigma X_2 & \cdots & \Sigma X_k \\
\Sigma X_2 & \Sigma X_2^2 & \cdots & \Sigma X_2X_k \\
\cdot & \cdot & \cdot & \cdot \\
\Sigma X_k & \Sigma X_2X_k & \cdots & \Sigma X_k^2
\end{vmatrix} \tag{20}
$$

If we use (19) and the relation $\Sigma(X_1 - \bar{X}_1)^2 = \Sigma X_1^2 - (\Sigma X_1)^2/N$ we can reduce (18) to the form *

$$r^2_{1\ 23\cdot\ \cdot k} =$$

$$\frac{a_1\Sigma X_1 + b_{12}\Sigma X_1X_2 + \cdots + b_{1k}\Sigma X_1X_k - (\Sigma X_1)^2/N}{\Sigma X_1^2 - (\Sigma X_1)^2/N} \qquad (21)$$

It can be shown that *the coefficient of multiple correlation between* X$_1$ *and* X$_2$, . . . , X$_k$ *is the same as the simple coefficient of correlation between* X$_1$ *and* X$_1'$, *the least squares linear estimate of* X$_1$ *from* X$_2$, . . . , X$_k$.

As an example, let us consider the data of section 20 and determine the coefficient of multiple regression between Y and X_1, X_2. Changing the notation in (21) slightly, we find

$$r^2_{Y\cdot X_1X_2} = \frac{b_0\Sigma Y + b_1\Sigma X_1Y + b_2\Sigma X_2Y - (\Sigma Y)^2/N}{\Sigma Y^2 - (\Sigma Y)^2/N}$$

$$= \frac{-5\ 9357\times15+1\ 2194\times71+1.7484\times32-(15)^2/5}{55 - (15)^2/5}$$

$$= 0.8491, \quad r_{Y\ X_1X_2} = 0.92$$

29. Partial correlation. Suppose that we have k variables $X_1, \ldots\ , X_k$ If we estimate X_1 from the others, excepting X_2, by means of the linear regression equation

$$X_1' = a_1 + b_{13}X_3 + b_{14}X_4 + \cdots + b_{1k}X_k$$

and likewise estimate X_2 from the others, omitting X_1, by means of the equation

$$X_2' = a_2 + b_{23}X_3 + b_{24}X_4 + \cdots + b_{2k}X_k$$

then the *coefficient of partial correlation* between X_1 and X_2 is the coefficient of correlation between the residuals $X_1 - X_1'$ and $X_2 - X_2'$. It may be regarded as the correlation between X_1 and X_2 when the effect of the remaining variables on each of them has been removed by linear regression equations.

* Another form is $r^2_{1.23\quad k} = 1 - R/R_{11}$, where R and R_{11} are defined in section 29.

If we estimate each of the variables X_1 and X_2 from all the others by means of the equations

$$X_1' = a_1 + b_{12}X_2 + b_{13}X_3 + \cdots + b_{1k}X_k$$

$$X_2' = a_2 + b_{21}X_1 + b_{23}X_3 + \cdots + b_{2k}X_k$$

the coefficient of partial correlation between X_1 and X_2 may also be defined as the square root of

$$r^2_{12\,3\quad k} = b_{12}b_{21} \tag{22}$$

The partial correlation coefficient can be simply expressed by means of determinants. Let

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{vmatrix} \tag{23}$$

in which r_{ij} is the coefficient of correlation between X_i and X_j (sometimes called the coefficient of *total correlation* between X_i and X_j). Then *

$$r_{12\,3\quad k} = -\frac{R_{12}}{(R_{11}R_{22})^{1/2}} \tag{24}$$

where R_{ij} is the cofactor of r_{ij} in the determinant R, that is, $(-1)^{i+j}$ times the determinant remaining after the ith column and jth row have been struck out of R.

The coefficient $r_{12\,3\quad k}$ is said to be of *order* $k - 2$; formula (24) expresses the coefficient of order $k - 2$ in terms of the coefficients r_{ij} of order zero.

We shall compute a coefficient of order one from the data of Table 8. Here we have $N = 5$, and replacing Y by X_3,

$$\Sigma X_1 = 18 \qquad \Sigma X_2 = 13 \qquad \Sigma X_3 = 15$$

$$\Sigma X_1^2 = 110 \qquad \Sigma X_2^2 = 45 \qquad \Sigma X_3^2 = 55$$

$$\Sigma X_1X_2 = 25 \qquad \Sigma X_1X_3 = 71 \qquad \Sigma X_2X_3 = 32$$

* For a proof of this statement and for a fuller discussion of partial and multiple correlation see H. L. Rietz, "Mathematical Statistics" (Carus Monograph 3), Open Court Publishing Co , Chicago, 1926.

$$r_{12} = \frac{\Sigma X_1X_2 - \Sigma X_1 \cdot \Sigma X_2/N}{[\Sigma X_1^2 - (\Sigma X_1)^2/N]^{\frac{1}{2}}[\Sigma X_2^2 - (\Sigma X_2)^2/N]^{\frac{1}{2}}}$$

$$= \frac{25 - 18 \cdot 13/5}{[110 - (18)^2/5]^{\frac{1}{2}}[45 - (13)^2/5]^{\frac{1}{2}}} = \frac{-21\ 8}{(45.2)^{\frac{1}{2}}(11.2)^{\frac{1}{2}}}$$

$$= -\ 0.9689$$

$$r_{13} = \frac{71 - 18 \cdot 15/5}{(45.2)^{\frac{1}{2}}[55 - (15)^2/5]^{\frac{1}{2}}} = \frac{17}{(45.2)^{\frac{1}{2}}(10)^{\frac{1}{2}}} = 0.7996$$

$$r_{23} = \frac{32 - 13 \cdot 15/5}{(11\ 2)^{\frac{1}{2}}(10)^{\frac{1}{2}}} = \frac{-7}{(11.2)^{\frac{1}{2}}\,(10)^{\frac{1}{2}}} = -\ 0.6614$$

The determinant (23) is

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}$$

of which the cofactors needed are

$$R_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}, \quad R_{12} = -\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}, \quad R_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix}$$

Formula (24) becomes

$$r_{12\ 3} = \frac{-R_{12}}{(R_{11}R_{22})^{\frac{1}{2}}} = \frac{r_{12} - r_{13}r_{23}}{(1 - r_{13}^2)^{\frac{1}{2}}(1 - r_{23}^2)^{\frac{1}{2}}}$$

$$= \frac{-0\ 9689 - 0.7996(-0.6614)}{[1 - (0.7996)^2]^{\frac{1}{2}}[1 - (-0.6614)^2]^{\frac{1}{2}}}$$

$$= \frac{-0.4400}{(0\ 360619 \times 0.5625+)^{\frac{1}{2}}} = \frac{-0.4400}{0.45038} = -\ 0.977$$

EXERCISES

1. Find the coefficient of correlation between I.Q. and score in reading vocabulary test in Table J, page 43

2. Find the coefficient of correlation between number of red blood cells and hemoglobin (*a*) for the men of Table O, (*b*) for the women of Table O. (Adapted from data quoted by Dunn, *Physiological Reviews*, vol. 9)

3. Table P gives the lengths of right chela (pincer) and of carapace (shell) of 470 females of a species of deep-water crab (Schuster, *Biometrika*, vol 2) (*a*) Find the coefficient of correlation between chela length and carapace length. (*b*) Find the equations of both lines of regression. (*c*) Find the correlation ratio of chela length on carapace length

TABLE O

NUMBER OF RED BLOOD CELLS AND AMOUNT OF HEMOGLOBIN, 20 MEN AND 12 WOMEN

Men				Women	
Red blood cells (mıl-lıons per cubıc mıllımeter)	Hemoglobın (grams per 100 cc)	Red blood cells (mıl-lions per cubıc mıllımeter)	Hemoglobın (grams per 100 cc)	Red blood cells (mıl-lıons per 100 cc.)	Hemoglobin (grams per 100 cc.)
4 27	14 00	4 93	16 20	3 89	12 12
4 40	14 41	4 97	15 40	3 95	12 10
4 52	14 02	5 00	16 40	3 97	11 90
4 56	14 20	5 02	15 52	4 15	13 20
4 58	14 50	5 15	16 50	4 20	13 10
4 64	14 30	5 20	15.75	4 26	13 50
4 72	14 70	5 36	16 10	4 31	13 40
4 80	15 10	5 49	16 70	4 38	14 80
4 84	15 00	5 57	17 17	4 40	13 50
4 89	15 60	5 62	16 61	4 45	13 88
				4 56	14 00
				4 72	14 60

TABLE P

CORRELATION TABLE OF LENGTHS OF RIGHT CHELA AND OF CARAPACE IN 470 CRABS

Length of carapace in mıllı-meters	Length of rıght chela ın mıllımeters										
	8–9	9–10	10–11	11–12	12–13	13–14	14–15	15–16	16–17	17–18	18–19
9 5–10 0											2
9 0– 9.5								3	4	4	2
8 5– 9 0				1			1	6	15	11	2
8 0– 8 5			1		2	1	15	51	25	3	
7 5– 8 0			4	1	3	9	50	34	4	2	
7 0– 7 5		1	2		2	43	43	2			
6 5– 7 0		1		1	23	28	3				
6 0– 6 5			1	20	8						
5 5– 6 0	1	2	10	9							
5 0– 5 5		2	3								
4 5– 5 0	1	6									
4 0– 4 5	2										

4. (*a*) Find the coefficient of correlation in Table Q (*b*) Find the correlation ratio η_{YX}. (*c*) Draw a diagram similar to Fig 8, page 53

TABLE Q

CORRELATION TABLE OF AMOUNT OF NITROGEN USED AS FERTILIZER AND YIELD OF WHEAT

Yield of wheat in bushels per acre, *Y*	Nitrogen applied in pounds per acre, *X*								
	0–20	20–40	40–60	60–80	80–100	100–120	120–140	140–160	160–180
32–36				6	15	10	4	6	2
28–32			1	18	20	9	5	1	
24–28		1	15	20	3				
20–24		2	12						
16–20		10	2						
12–16		8							
8–12	4	4							
4–8	10								
0–4	6								

5. (*a*) From the data of Table N find the coefficient of multiple correlation between grade-point average and the other variables (*b*) Find the coefficient of partial correlation between grade-point average and reading comprehension. (*c*) Find the coefficient of partial correlation between grade-point average and reading rate.

6. Using the data of Table N, but leaving grade-point average entirely out of consideration, find the coefficient of partial correlation between (*a*) reading comprehension and reading rate, (*b*) intelligence and reading comprehension, (*c*) intelligence and reading rate.

7. Tables R_1, R_2, R_3 are correlation tables of age and weight, height and weight, age and height, of 1138 boys. (Data were selected by Tippett's Random Sampling Numbers from more extensive data of Isserlis, *Biometrika*, vol 11.) (*a*) Find the coefficient of multiple correlation of weight on height and age. Find the coefficient of partial correlation between (*b*) weight and height, (*c*) weight and age, (*d*) height and age

8. (*a*) Find the correlation ratio of weight on height in Table R_2 (*b*) Fit a parabola of least squares to the means of columns of Table R_2, weighting each mean by the number of individuals in the corresponding column

TABLE R_1

Correlation Table of Ages and Weights of 1138 Boys

Weight in pounds	Age in years									
	5	6	7	8	9	10	11	12	13	14
31	2	4	2			1				
36	6	13	16	5		4	2			
41	10	25	47	24	8	10	1			
46	17	49	33	51	34	37	2	3	1	
51	3	27	22	52	38	37	10	5		
56		6	13	16	36	30	32	22	6	1
61			2	4	17	12	26	24	17	2
66				1	3	6	32	25	17	7
71				1	3	1	15	17	26	13
76					1		5	11	29	12
81								5	11	2
86							3	4	5	3
91								2	4	1
96									3	1
101									1	1

TABLE R_2

Correlation Table of Heights and Weights of 1138 Boys

Weight in pounds	Height in inches.										
	34	37	40	43	46	49	52	55	58	61	64
31	2	6	1								
36		24	16	6							
41		6	50	59	10						
46			51	74	73	21	8				
51	1		2	47	91	48	4		1		
56				7	39	76	37	2			1
61					2	44	49	9			
66						12	64	14	1		
71						3	32	40	1		
76							14	35	9		
81							2	5	11		
86								7	8		
91								1	3	3	
96									1	3	
101										1	1

TABLE R_3

Correlation Table of Ages and Heights of 1138 Boys

Height in inches	Age in years									
	5	6	7	8	9	10	11	12	13	14
34	2						1			
37	7	16	5	5		1	2			
40	22	50	29	9	7	2	1			
43	7	46	61	50	19	7	3			
46	.	12	35	56	58	35	9	8	2	
49			4	22	39	54	47	26	12	
52			1	11	15	33	45	52	41	12
55					2	5	18	26	43	19
58				1			2	6	16	10
61							.		5	2
64						1			1	

CHAPTER V

THE BINOMIAL AND NORMAL DISTRIBUTIONS

30. Binomial distribution. In textbooks on college algebra * it is shown that if p is the probability that an event will happen, and $q(=1-p)$ is the probability that it will fail to happen, in a single trial, then the probability that it will happen exactly X times in N trials is

$$C_X^N p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \tag{1}$$

C_X^N being the number of combinations of N things taken X at a time. For example, in tossing a die the probability of throwing an ace is $\frac{1}{6}$, that is, $p = \frac{1}{6}$, $q = \frac{5}{6}$. In tossing a die 5 times (or 5 dice once) the probability of throwing exactly 3 aces is

$$\frac{5!}{3!\,2!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{250}{6^5} = 0.0321$$

If a long series of hospital records shows that 40 per cent of cases of a certain disease fail to recover, $p = 0.4$ may be regarded as the probability of dying from the disease. Then, just as in the dice problem above, the probability that exactly 3 patients out of a given set of 5 will die is

$$\frac{5!}{3!\,2!}(0.4)^3(0.6)^2 = 0.2304$$

The expression (1) is the general term in the binomial expansion of $(q + p)^N$. Thus, the successive terms of this expansion give the probabilities that the event will happen not at all, once, twice, . . . , N times, respectively. If we compute all these

* See, for example, H L. Rietz and A. R. Crathorne, "College Algebra," Henry Holt & Co , New York.

probabilities for the binomial $(0.6 + 0\ 4)^5$ we have the results shown in Table 12.

TABLE 12

PROBABILITY OF X DEATHS IN 5 CASES OF A DISEASE FOR WHICH THE MORTALITY RATE IS 40 PER CENT

X	Probability
0	0 07776
1	0 25920
2	0 34560
3	0 23040
4	0 07680
5	0 01024
	1 00000

If a new treatment for this disease were being tried out, 5 patients would be too small a number from which to form a judgment as to the efficacy of the treatment, but suppose it had been administered to 10 patients and all but 1 of them recovered. The probability of no deaths in a group of 10 patients is

$$P(0) = (0.6)^{10} = 0.00605-$$

The probability of 1 death is

$$P(1) = 10(0.4)(0\ 6)^9 = 0.04031$$

The probability of fewer than 2 deaths is

$$P(<2) = P(0) + P(1) = 0.04636$$

In sets of 10 patients we should then expect at most 1 death 4.6 per cent of the time, or about once in 20 times, so we should be somewhat doubtful that the treatment is effective. Our doubt will be strengthened if we consider the probability of a deviation in either direction from the *expected value,* which is 40 per cent of 10, or 4. We have already found that the probability of a deviation of 3 or more below the expected value is $P(\leqq 1) = P(<2) = 0.04636$. Similarly we can find that the probability of a deviation of 3 or more above the expected value is $P(\geqq 7) = P(>6) = 0.05476$. Thus the probability of a numerical deviation of 3 or more from expectation is $P\ (\,|\,X - 4\,|\, \geqq 3) = 0.04636 + 0.05476 =$ C.10112. That is, as a matter of pure chance we should find a deviation of 3 or more deaths from the expected number oftener than once in 10 times.

31. Normal distribution. As the number of cases grows larger the computation of the various probabilities becomes very tedious. Fortunately, however, as the number of cases increases the binomial distribution approaches the so-called *normal distribution*

$$YdX = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-(X-\mu)^2/2\sigma^2} dX \tag{2}$$

In section 5 was given the transformation that would change the normal distribution (2) into the simpler form

$$Ydx = \varphi(x)dx, \quad \varphi(x) = (2\pi)^{-\frac{1}{2}}e^{-x^2/2} \tag{3}$$

Values of $\varphi(x)$, that is, the ordinates of the normal curve, are given in Table I at the end of this book.

The function

$$\varphi_{-1}(x) = \int_{-\infty}^{x} \varphi(x)dx = \int_{-\infty}^{x} (2\pi)^{-\frac{1}{2}}e^{-x^2/2}dx \tag{4}$$

gives the probability that a normal variable, with mean zero and standard deviation unity, will be less than x. It is the area under the normal curve, from the extreme left to the ordinate corresponding to the abscissa x. The difference $\varphi_{-1}(x_2) - \varphi_{-1}(x_1)$ gives the probability that a random variate * will fall between x_1 and x_2; it is the area under the curve between $x = x_1$ and $x = x_2(x_2 > x_1)$.

It has seemed more desirable in this book to deal with the probability that a normal deviate will be greater than x. This probability will also be found in Table I. If we denote it by $P(>x)$, then we have

$$P(>x) = \int_{x}^{\infty} \varphi(x)dx = 1 - \varphi_{-1}(x)] \tag{5}$$

Thus, $P(>x)$ is the area under the normal curve to the right of the ordinate corresponding to x. Further, by means of (5) it can be shown that the probability that x will be between x_1 and x_2 is $P(>x_1) - P(>x_2)$, which is the area under the curve between x_1 and $x_2(x_2 > x_1)$. This may be expressed by the formula

$$\int_{x_1}^{x_2} \varphi(x)dx = P(>x_1) - P(>x_2) \tag{6}$$

To illustrate the way in which a binomial distribution approaches the corresponding normal distribution, we shall fit a normal distribution to the binomial $(0.6 + 0.4)^{25}$ which gives, for

* A *variate* may be defined as a particular value of a variable.

example, the probabilities of various numbers of deaths from 0 to 25, in 25 cases of a disease for which the mortality rate is 40 per cent.

It can easily be shown * that for the binomial distribution $(q + p)^N$ the mean number of happenings is Np and that the standard deviation of the number of happenings is $(Npq)^{1/2}$ It can also be shown, by integration, that the mean of the normal distribution (2) is μ, that its standard deviation is σ, and that the area underneath the curve is unity † The method of fitting a normal distribution to the binomial consists in choosing the mean and the standard deviation of the normal curve equal respectively to the corresponding parameters of the binomial distribution.

In the present case we have $Np = 25 \times 0.4 = 10$, which is the mean or expected number of deaths in a group of 25 patients, also $\sigma = (Npq)^{1/2} = (25 \times 0.4 \times 0.6)^{1/2} = 6^{1/2} = 2.44949$. Consequently the corresponding normal distribution is

$$YdX = \frac{1}{2.45(2\pi)^{1/2}} e^{-(X-10)^2/2\times 6} dX \tag{7}$$

The two distributions may be compared graphically by drawing the curve

$$Y = \frac{1}{2.45(2\pi)^{1/2}} e^{-(X-10)^2/2\times 6} \tag{8}$$

and, on the same set of axes, erecting ordinates equal to the binomial probabilities. Although this is unnecessary, it will be done here to give a clearer insight into the meaning of the normal approximation to the binomial distribution.

* See Yule and Kendall, "An Introduction to the Theory of Statistics," Charles Griffin & Co., Ltd , London, 1937.

† Another parameter connected with the normal curve is the *probable error* or *probable deviation* It is defined as that deviation which, taken on each side of the mean, will include half of the area under the normal curve. If an item is selected at random from a normal distribution, it is equally probable that its absolute deviation from the mean will be greater than one probable error or less than that value. The probable error may be found by the approximate formula P.E. = 0.67449σ.

In order to use available tables we set $(X - 10)/6^{1/2} = x$, then (8) assumes the form

$$Y = \frac{1}{2.45(2\pi)^{1/2}} e^{-x^2/2} = \frac{1}{2.45}\varphi(x) \tag{9}$$

From Table I we can obtain the values necessary for plotting the curve. The binomial probabilities, worked out by the formula

$$P(X) = C_X^{25}(0.4)^X(0.6)^{25-X} = \frac{25!}{X!(25 - X)!}(0.4)^X(0.6)^{25-X} \tag{10}$$

are shown in Table 13. Figure 9 shows the normal curve (9) with the probabilities (10) plotted as small circles.

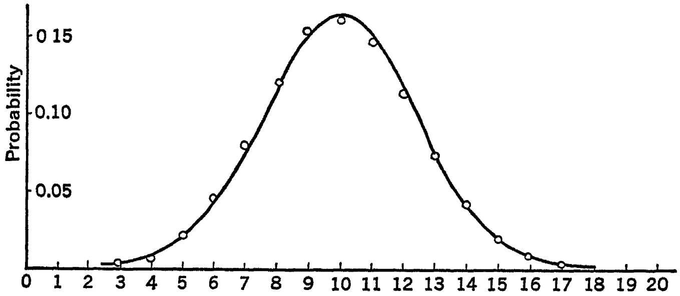


FIG 9—Normal Curve Fitted to a Binomial Distribution

To get a normal probability, we must find the area under the appropriate part of the normal curve. It must be remembered that the binomial distribution is discrete, in fact it is often called the *point binomial.* If then, for example, we wish to find the probability of 13 deaths by using the normal approximation we find the area under the normal curve from $X_1 = 12.5$ to $X_2 = 13.5$. This is given by the definite integral

$$\int_{12.5}^{13.5} \frac{1}{6^{1/2}(2\pi)^{1/2}} e^{-(X-10)^2/2\times 6}dX \tag{11}$$

We make the transformation

$$\frac{X - 10}{6^{1/2}} = x, \quad \frac{dX}{6^{1/2}} = dx$$

which gives

$$x_1 = \frac{12.5 - 10}{6^{1/2}} = 1.021, \quad x_2 = \frac{13\ 5 - 10}{6^{1/2}} = 1.429$$

and changes the integral (11) into

$$\int_{1\ 021}^{1\ 429} (2\pi)^{-1/2} e^{-x^2/2} dx = \int_{1\ 021}^{1\ 429} \varphi(x) dx$$

$$= P(>1\ 021) - P(>1.429)$$

$$= 0.1536 - 0.0763 = 0.0773$$

from Table I. The true probability, found by using the binomial formula (10), is, to six decimal places, 0 075967.

One advantage of the normal distribution is illustrated by the following examples: The probability of 14 or more deaths is given by

$$P\left(> \frac{13\ 5 - 10}{6^{1/2}}\right) = P(>1.429) = 0.0763$$

The probability of between 7 and 13 deaths, inclusive, is

$$P\left(>\frac{6\ 5-10}{6^{1/2}}\right) - P\left(> \frac{13\ 5-10}{6^{1/2}}\right) = P(>-1.429) - P(>1.429)$$

$$= 1 - 2P(>1.429) = 1 - 2 \times .0763 = 0.8474$$

(It can easily be shown that $P(> - x_1) = 1 - P(>x_1)$.)

If these probabilities were calculated by the exact binomial formula it would be necessary to calculate all the individual probabilities included and then to sum them.

The normal approximation * to the binomial $(0.6 + 0.4)^{25}$ is shown in Table 13 It is seen that the approximation is only fairly good. It is better near the expected value $\bar{X} = 10$ than it is in the tails of the distribution. It would have been better for larger values of N(say $N = 100$, in which case it would have been too laborious to compute the binomial probabilities); it would have been worse for p not so close to ½. It is difficult to lay down

* Obtained by using a seven-place table (Table II of Karl Pearson's "Tables for Statisticians and Biometricians," Part 1).

fixed rules about when the normal distribution may be used as an approximation to the binomial; this depends upon the accuracy of the results desired, and good judgment in the matter comes only with experience.*

TABLE 13

PROBABILITY OF X DEATHS IN 25 CASES OF A DISEASE FOR WHICH THE MORTALITY RATE IS 40 PER CENT

X	Probability	Normal approximation
0	0.000 003	0 000 044
1	0 000 047	0 000 208
2	0 000 379	0 000 840
3	0 001 937	0 002 882
4	0 007 104	0 008 389
5	0 019 891	0 020 726
6	0 044 203	0 043 419
7	0 079 986	0 077 206
8	0 119 980	0 116 415 −
9	0 151 086	0 149 001
10	0 161 158	0 161 725 −
11	0 146 507	0 149 001
12	0 113 950 +	0 116 415 −
13	0 075 967	0 077 206
14	0 043 410	0 043 419
15	0 021 222	0 020 726
16	0 008 843	0 008 389
17	0 003 121	0 002 882
18	0 000 925 −	0 000 840
19	0 000 227	0 000 208
20	0 000 045 +	0 000 044
21	0 000 007	0 000 008
22	0 000 001	0 000 001
23		
24		
25		
Total	0 999 999	0 999 994

If a better approximation is desired we can use a Gram-Charlier type A distribution. (See section 33.) For extremely small values of p combined with large values of N the Poisson exponential function may sometimes be used as an approximation.†

32. Fitting a normal distribution to observed data. The normal distribution is not limited in its uses to approximating the binomial, for experience has shown that many sets of data are distributed approximately like the normal distribution. Thus, suppose that we know that the heights of a group of men are normally distributed, with mean 68 inches and standard deviation

* For a discussion of this question and of other approximations to the binomial distribution, see Burton H. Camp, "Probability Integrals for the Point Binomial," *Biometrika*, vol 16, 1924, pp 163–171; and Camp's book, "The Mathematical Part of Elementary Statistics," D C. Heath & Co , Boston, 1931.

† See Thornton C. Fry, "Probability and Its Engineering Uses," D. Van Nostrand Co., New York, 1928.

2 inches, and wish to find the probability that the height of a man taken at random from the group is between 66 inches and 69 inches. We calculate

$$x_1 = \frac{66 - 68}{2} = -1, \quad x_2 = \frac{69 - 68}{2} = 0.5$$

$$P(>-1) - P(>0.5) = 1 - P(>1) - P(>0.5)$$
$$= 1 - 0.1587 - 0.3085 = 0.5328$$

which is the required probability. If 100 men were taken at random from such a population, we should expect to find about 100 × 0.5328, or 53 of them, between 66 and 69 inches tall.

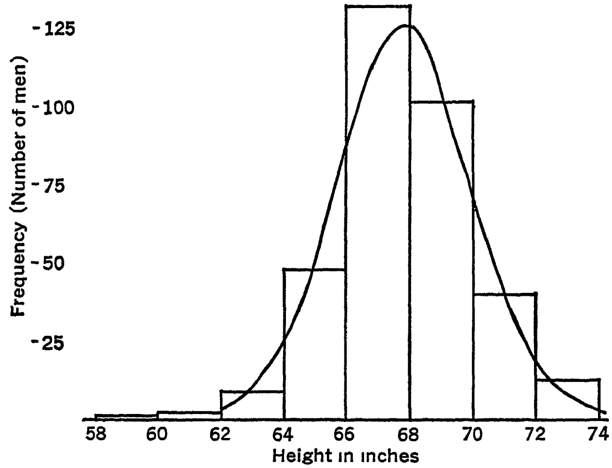


FIG 10—Normal Curve Fitted to Distribution of Heights of Men.

Such data as the above may be graduated by means of the normal distribution, that is, a normal distribution may be fitted to them by making its mean coincide with their mean and its standard deviation equal to their standard deviation. We shall illustrate the process by fitting a normal curve to the distribution of heights of men given in Table 1.

The mean of this distribution was found to be 67.84 inches; the standard deviation, after Sheppard's correction had been applied to the variance, was 2 17 inches. The equation of the corresponding normal curve is

$$YdX = \frac{1}{2.17(2\pi)^{1/2}} \exp\left[-\frac{(X-67.84)^2}{2(2.17)^2}\right] dX \qquad (12)$$

where X is to be expressed in inches. The curve is graphed in conjunction with the histogram of heights in Fig 10. In constructing the graph, we must multiply each value of Y by $N(=346)$.

To obtain the theoretical frequencies in the various classes, we express the class limits as deviations from the mean, divided by the standard deviation. Obviously the theoretical proportional frequencies must be multiplied by the total frequency N. The work is shown in Table 14. For purposes of comparison the observed frequencies are shown in the last column.

TABLE 14

Normal Distribution Fitted to the Frequency Distribution of Heights of a Group of Men

Class limit X	$x = \frac{X - 67\ 84}{2\ 17}$	$P(> x)$	Difference	346 × difference = theoretical frequency	Observed frequency
58	−4 53	1 0000			
			0 0002	0.1	1
60	−3 61	0 9998			
			0 0034	1 2	2
62	−2 69	0 9964			
			0 0348	12 0	9
64	−1 77	0 9616			
			0 1593	55 1	48
66	−0 85	0 8023			
			0 3302	114 2	131
68	0 07	0 4721			
			0 3134	108 4	102
70	1 00	0 1587			
			0.1313	45 4	40
72	1 92	0 0274			
			0 0251	8 7	13
74	2 84	0 0023			
			0 0022	0 8	
76	3 76	0 0001			
Total			0 9999	345 9	346

33. Gram-Charlier type A distribution. If a better approximation to a binomial distribution is desired we can, as suggested in section 31, use a Gram-Charlier type A distribution, which also provides a better fit than the normal function to an observed frequency distribution such as the distribution of heights considered above. This distribution consists of an expansion in terms of the normal function and its derivatives, viz ,

$$\begin{aligned} YdX &= Y\sigma\, dx \\ &= [a_0\varphi(x) + a_1\varphi_1(x) + a_2\varphi_2(x) + a_3\varphi_3(x) + \cdots]dx \qquad (13) \end{aligned}$$

where $\varphi_k(x)$ is the kth derivative of $\varphi(x) = (2\pi)^{-\frac{1}{2}}e^{-x^2/2}$, and and $x = (X - \bar{X})/\sigma$, $dx = dX/\sigma$. It can be shown that if we use x instead of X, that is, if we express our variable as a deviation from the mean divided by the standard deviation, then $a_0 = 1$, $a_1 = a_2 = 0$, and through the term in $\varphi_4(x)$, which is as far as we should often want to carry the expansion,

$$YdX = Y\sigma dx = \left[\varphi(x) - \frac{1}{6}\frac{\mu_3}{\sigma^3}\varphi_3(x) + \frac{1}{24}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi_4(x)\right]dx \quad (14)$$

The area under any portion of the Gram-Charlier curve

$$Y = \frac{1}{\sigma}\left[\varphi(x) - \frac{1}{6}\frac{\mu_3}{\sigma^3}\varphi_3(x) + \frac{1}{24}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi_4(x)\right] \qquad (15)$$

is given by

$$\int YdX = \int Y\sigma dx = \varphi_{-1}(x) - \frac{1}{6}\frac{\mu_3}{\sigma^3}\varphi_2(x) + \frac{1}{24}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi_3(x) \quad (16)$$

between the appropriate limits.

For the binomial distribution we have *

$$\mu_2 = Npq,\ \mu_3 = Npq(q - p),\ \mu_4 = Npq(1 - 6pq) + 3N^2p^2q^2 \quad (17)$$

Consequently (16) reduces to

$$\int YdX = \varphi_{-1}(x) - \frac{q - p}{6(Npq)^{\frac{1}{2}}}\varphi_2(x) + \frac{1 - 6pq}{24Npq}\varphi_3(x) \qquad (18)$$

For a further discussion of the Gram-Charlier distribution

* Thornton C Fry, "Probability and Its Engineering Uses," D Van Nostrand Co , New York, 1928

and examples of fitting observed data and the point binomial by means of it, the student is referred to Fry,* Camp,† and Fisher.‡

34. Testing the significance of a mean when the population standard deviation is known. If the population from which we are taking samples is normal, with mean μ and variance σ^2, then, as has long been well known, the means of samples of size N will themselves be normally distributed, with mean μ and variance σ^2/N (i.e., standard deviation $\sigma/N^{1/2}$).§ This fact enables us to use the normal distribution to test the significance of a mean, provided we know the standard deviation of the population.

Suppose, for example, that we know from considerable experience that the average breaking strength of a certain type of cotton thread is 7.50 ounces, and that the standard deviation is 1.20 ounces. A sample of 9 pieces of thread shows a mean breaking strength of 6.52 ounces. If we assume that the population of breaking strengths is normally distributed we can find the probability that the consignment from which the sample was taken is below standard. The standard deviation of the mean of a sample of 9 is

$$\sigma_{\bar{X}} = \frac{\sigma}{N^{1/2}} = \frac{1.20}{3} = 0.40 \text{ oz.}$$

We now calculate

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{6.52 - 7\ 50}{0.40} = -\ 2.45$$

The probability that the mean of a sample of 9 will be this far or farther below the population mean is

$$\varphi_{-1}(-2.45) = P(>2\ 45) = 0\ 0071$$

That is, in samples of 9, such a deviation below the general average would be expected only about 7 times in 1000, and we conclude that the product is inferior.

* Thornton C. Fry, "Probability and Its Engineering Uses," D. Van Nostrand Co , New York, 1928.

† Burton H Camp, "The Mathematical Part of Elementary Statistics," D C Heath & Co , Boston, 1931

‡ Arne Fisher, "The Mathematical Theory of Probabilities," The Macmillan Co , New York, 1930.

§ Even if the population is not normal, the distribution of means will usually be approximately normal

It should perhaps be emphasized that "significance" is a relative term. Thus, one person might regard a deviation as significant if the probability of the occurrence of a greater deviation were 0 05. Another might regard it as significant only if this probability were 0.001. It is largely a subjective matter and depends upon the chances that the individual is willing to take that his judgment may be wrong. Many investigators are willing to regard as *significant* any deviation (or difference) for which the probability of a greater deviation is 0.05, and as *highly significant* any deviation for which this probability is 0.01 or less.

It is conceivable that we might know the population standard deviation but not the population mean. For instance, we might be prepared to believe that the variability of strength of a certain type of thread is about the same although produced by different factories, but that the average value differed from factory to factory. Suppose that in the above illustration the 9 pieces of thread are from a certain factory and that we wish to test the hypothesis that the mean breaking strength of this type of thread produced by the factory is not below 6 ounces, on the assumption that $\sigma = 1.20$ ounces. We calculate

$$\frac{6.52 - 6}{1.20/9^{1/2}} = 1.30, \quad P(>1.30) = 0.0968$$

so that if the mean breaking strength of the product put out by this factory were 6 ounces, we should, in samples of 9, find a mean breaking strength of 6.52 ounces or greater nearly once in 10 times. The mean breaking strength of the product of this factory could very easily be as low as 6 ounces and still yield a sample of 9 having a mean breaking strength of 6.52 ounces.

35. Fiducial or confidence limits. Suppose that, with a known population standard deviation of 1.20 ounces, we want to set limits, as judged from our sample mean of 6.52 ounces, within which we should have some confidence that the population means lies. If we take the probability

$$P\left(>\frac{|\bar{X} - \mu|}{\sigma/N^{1/2}}\right) = P\left(>\frac{|6.52 - \mu|}{0\ 40}\right)$$

and set it equal to 0.02, say, we find * from Table II

$$|\,6.52 - \mu\,|\,/0.40 = 2.33, \quad |\,6.52 - \mu\,| = 0\,93$$

$$\mu = 6.52 \pm 0.93, \quad \mu_1 = 5.59, \quad \mu_2 = 7.45$$

The value $\mu_2 = 7\,45$ may be called the 1 per cent *fiducial value* of μ for $\bar{X} = 6\,52$ and $\sigma = 1.20$. Sımılarly, $\mu_1 = 5.59$ may be termed the 99 per cent fiducıal value of μ correspondıng to the given values of $\bar{X}$ and σ. The two values may be called the 98 per cent fiducial or *confidence lımıts* for μ correspondıng to the given $\bar{X}$ and σ. For suppose that from a sample mean $\bar{X}$ we should assert that the populatıon mean μ ıs between the limıts $\bar{X} - 2.33\sigma/N^{1/2}$ and $\bar{X} + 2.33\sigma/N^{1/2}$, that is,

$$\bar{X} - \frac{2\,33\sigma}{N^{1/2}} < \mu < \bar{X} + \frac{2.33\sigma}{N^{1/2}}$$

This inequality ıs equivalent to

$$\mu - \frac{2.33\sigma}{N^{1/2}} < \bar{X} < \mu + \frac{2.33\sigma}{N^{1/2}}$$

But the probabılıty that $\bar{X}$ wıll satisfy this last ınequalıty is 0.98, so that in the long run we should be rıght in our assertıon regardıng the populatıon mean 98 tımes out of 100 and wrong twice out of 100 tımes—once because the observed $\bar{X}$ fell below $\mu - 2.33/N^{1/2}$ and once because it was above $\mu + 2\,33/N^{1/2}$, so that the inversıon of the second ınequality above supplıed an inequalıty which failed to include μ in these ınstances

36. Testing the significance of the difference between two means when the population standard deviation is known. If the varıables X_1 and X_2 are normally distrıbuted with means μ_1 and μ_2 respectıvely, theır dıfference $X_1 - X_2$ wıll be normally dıstrıbuted with mean $\mu_1 - \mu_2$.

The variance of the dıfference between two variables is the sum of their variances dimınished by twice their covarıance. In symbols this may be stated

$$\sigma^2_{X_1-X_2} = \sigma^2_{X_1} + \sigma^2_{X_2} - 2r\sigma_{X_1}\sigma_{X_2} \tag{19}$$

* Note that the probabılıty of an absolute or numerıcal deviation is twıce that of the correspondıng algebraıc devıatıon, and that consequently we find in Table II the value of x coıresponding to $P(> x) = 0\,01$.

In the last term, r is the coefficient of correlation between the variables X_1 and X_2, and $r\sigma_{X_1}\sigma_{X_2}$ is their covariance. If X_1 and X_2 are uncorrelated this last term drops out and the foregoing formula reduces to

$$\sigma^2_{X_1-X_2} = \sigma^2_{X_1} + \sigma^2_{X_2} \tag{20}$$

The variance of the difference between two uncorrelated means is thus

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma^2_{X_1}}{N_1} + \frac{\sigma^2_{X_2}}{N_2} \tag{21}$$

where N_1 and N_2 are the respective numbers in the samples from which the means $\bar{X}_1$ and $\bar{X}_2$ were calculated.

It may sometimes happen that it is reasonable to suppose that the populations from which the samples have been drawn have the same variance σ^2. In this case, (21) assumes the still simpler form

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right). \tag{22}$$

For example, experience might show that the standard deviation of the breaking strength of a type of thread such as that considered earlier in the chapter could be regarded as 1.20 ounces. Suppose that under this assumption a sample of 9 pieces from one factory showed a mean strength of 6.52 ounces, while a sample of 16 pieces from another factory showed a mean strength of 7.20 ounces. We wish to know whether we can regard the general average output of the second factory as superior. We can test the hypothesis that the two samples were drawn from populations with the same mean, that is, $\mu_1 = \mu_2$, or that the difference $\bar{X}_1 - \bar{X}_2$ has the mean $\mu_1 - \mu_2 = 0$. Here we have

$$\sigma_{\bar{X}_1-\bar{X}_2} = 1.20\ (\tfrac{1}{9} + \tfrac{1}{16})^{1/2} = 1.20 \times \tfrac{5}{12} = 0.5$$

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1-\bar{X}_2}} = \frac{6.52 - 7.20}{0.5} = -\ 1.36$$

Actually we are concerned only with the absolute value of the difference. We find from tables of the normal distribution that the probability of a deviation greater than 1.36 above or below the mean is

$$2P(>1.36) = 2 \times 0.0869 = 0.1738$$

This would not disprove the hypothesis, and we can not conclude that the product of the second factory is, on the average, superior.

37. Testing the significance of the difference between two proportions. If, from a population in which the proportion of individuals possessing a certain characteristic is p, we take a large number of samples of size N, then the number X of individuals possessing this characteristic might range in these samples from 0 to N, but would on the average be $\bar{X} = Np$, and the standard deviation of this number would be $\sigma_X = (Npq)^{\frac{1}{2}}$, as was stated in section 31. The proportion of such individuals might have the values $0, 1/N, 2/N, \ldots, (N-1)/N, 1$, but would on the average have the value $\bar{p} = p$, and the standard deviation $\sigma_X/N = (pq/N)^{\frac{1}{2}}$.

If we draw repeatedly two samples, one of size N_1 and the other of size N_2, from a common population, and if we average the obtained values of p_1, and likewise those of p_2, we expect to find the average proportions to be $\bar{p}_1 = \bar{p}_2 = p$, the true population proportion, and the standard deviations to be

$$\sigma_{p_1} = \left(\frac{pq}{N_1}\right)^{\frac{1}{2}}, \quad \sigma_{p_2} = \left(\frac{pq}{N_2}\right)^{\frac{1}{2}} \tag{23}$$

respectively.

Suppose we wish to test the hypothesis that two observed proportions, p_1 and p_2, obtained from samples of size N_1 and N_2 respectively, are consistent with sampling from a common population. On this hypothesis, the expected value of the difference $p_1 - p_2$ is $\bar{p}_1 - \bar{p}_2 = 0$, so that we wish to test whether the observed difference is significantly different from zero. The standard deviation of the difference, assuming no correlation between the proportions,* is

$$\sigma_{p_1-p_2} = (\sigma_{p_1}^2+\sigma_{p_2}^2)^{\frac{1}{2}} = \left(\frac{pq}{N_1} + \frac{pq}{N_2}\right)^{\frac{1}{2}} = (pq)^{\frac{1}{2}}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{\frac{1}{2}} \tag{24}$$

* The proportions might, for example, be correlated if the individuals in the first sample were brothers of those in the second

Since $p = \bar{p}_1 = \bar{p}_2$ is unknown, it must be estimated in order to evaluate (24), and theoretical considerations show that a good estimate based on the samples is the weighted mean of the two observed proportions, viz.,

$$p' = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

which is the proportion observed in the combined samples. If the numbers N_1 and N_2 are large, we may assume that the quantity

$$x = \frac{p_1 - p_2}{(p'q')^{1/2}(1/N_1 + 1/N_2)^{1/2}} \tag{25}$$

$(q' = 1 - p')$ is normally distributed, and so test the significance of the observed difference.

Suppose that in a group of 33 light-haired persons there are 26 with blue eyes, while in a group of 27 dark-haired persons there are only 9 with blue eyes. (See Table 15.) The proportion of blue-eyed persons in the light-haired group is $p_1 = {}^{26}\!/_{33} = 0.\dot{7}\dot{8}$, while in the dark-haired group it is $p_2 = {}^{9}\!/_{27} = {}^{1}\!/_{3} = 0.\dot{3}$. For the total group, $p' = {}^{35}\!/_{60} = {}^{7}\!/_{12} = 0.58\dot{3}$, $q' = 1 - {}^{35}\!/_{60} = {}^{25}\!/_{60} = {}^{5}\!/_{12} = 0.41\dot{6}$. Further, $N_1 = 33$, $N_2 = 27$, and from (25) we have

TABLE 15

HAIR-COLOR AND EYE-COLOR OF 60 PERSONS

	Light-haired	Dark-haired	Total
Blue-eyed.	26	9	35
Brown-eyed	7	18	25
Total	33	27	60

$$x = \frac{0.7879 - 0.3333}{(\frac{35}{60} \times \frac{25}{60})^{1/2}(\frac{1}{33} + \frac{1}{27})^{1/2}} = \frac{0.4546}{0.1279} = 3.55+$$

The probability of a greater numerical difference is

$$2P(>3.55) = 0.0004$$

It should be added that the test regarding the difference between the proportion of light-haired persons among those having blue eyes and among those having brown eyes yields precisely the same result. This is always true in a table such as the foregoing. We have for this test

$$p_1 = \tfrac{26}{35} = 0.7429, \quad p_2 = \tfrac{7}{25} = 0.2800, \quad N_1 = 35, \quad N_2 = 25$$

$$p' = \tfrac{33}{60}, \qquad q' = \tfrac{27}{60}, \qquad N = 60$$

$$x = \frac{0\ 7429 - 0\ 2800}{(\frac{33}{60} \times \frac{27}{60})^{\frac{1}{2}}(\frac{1}{35} + \frac{1}{25})^{\frac{1}{2}}} = \frac{0.4629}{0.1303} = 3.55+$$

38. Testing the significance of a correlation coefficient.[*] The significance of a correlation coefficient r may be tested by making the transformation

$$X = \frac{1}{2}\log_e\frac{1+r}{1-r} = 1.1513 \log_{10}\frac{1+r}{1-r} \tag{26}$$

Then X will be approximately normally distributed with standard deviation $1/(N-3)^{\frac{1}{2}}$, where N is the number of pairs of variates from which r was calculated. If ρ is the coefficient of correlation in the population from which the sample was drawn, we can also make the transformation

$$\mu = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} = 1.1513 \log_{10}\frac{1+\rho}{1-\rho} \tag{27}$$

and test whether X deviates significantly from μ.

Suppose that the scores in two tests administered to the same set of 20 students have a correlation of 0.65. Could we expect these same tests in general to yield a correlation of as much as 0.50?

We test the hypothesis that the population correlation is 0.50. That is, we assume that the population correlation is 0.50 and see whether the value 0 65 is unusual in such a case. If it is, we should reject the hypothesis. We find

$$X = \frac{1}{2}\log_e\frac{1+0.65}{1-0.65} = 0.77530$$

[*] See R A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol 1, 1921, part 4, pp 1–32.

$$\mu = \frac{1}{2}\log_e \frac{1 + 0.50}{1 - 0.50} = 0.54930$$

$$\sigma_X = \frac{1}{(N - 3)^{1/2}} = \frac{1}{(17)^{1/2}}$$

$$\frac{X - \mu}{\sigma_X} = (0.77530 - 0.54930)(17)^{1/2} = 0.9318$$

As a normal deviate this is, of course, not significant.

To test whether this correlation coefficient is significantly different from zero we calculate

$$\frac{X - 0}{\sigma_X} = 0.77530(17)^{1/2} = 3.197$$

The probability of a numerical deviation of this much or more is

$$2 \times P(>3\ 197) = 2 \times 0.0007 = 0\ 0014$$

which is extremely small, so that the observed correlation is decidedly significant.

An exact method of testing the significance of a correlation when the correlation in the population is zero will be given in Chapter VI.

In testing the significance of a partial correlation coefficient we proceed as above, except that the standard deviation of X is

$$\sigma_X = \frac{1}{(N - m - 3)^{1/2}} \tag{28}$$

where m is the number of variables eliminated. Thus in testing the significance of a partial coefficient $r_{123 \cdot\ k}$ we should use $\sigma\ \ = 1/(N - k - 1)^{1/2}$, since $k - 2$ variables have been eliminated.

39. Testing the significance of the difference between two correlation coefficients. To test the significance of the difference between two correlation coefficients r_1 and r_2, calculated from samples of N_1 and N_2 respectively, we make the logarithmic transformation

$$X_1 = \frac{1}{2}\log_e \frac{1 + r_1}{1 - r_1}, \quad X_2 = \frac{1}{2}\log_e \frac{1 + r_2}{1 - r_2} \tag{29}$$

Since

$$\sigma_{X_1} = \frac{1}{(N_1 - 3)^{1/2}}, \qquad \sigma_{X_2} = \frac{1}{(N_2 - 3)^{1/2}} \tag{30}$$

the standard deviation of the difference $X_1 - X_2$ is

$$\sigma_{X_1 - X_2} = (\sigma^2_{X_1} + \sigma^2_{X_2})^{1/2} = \left(\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}\right)^{1/2} \tag{31}$$

We therefore set

$$x = (X_1 - X_2) \div \left(\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}\right)^{1/2} \tag{32}$$

and regard x as a normal deviate, with unit standard deviation. We test whether x is significantly different from its expected value, zero.

Suppose that a correlation coefficient of $r_1 = 0.60$ has been obtained from a sample of size 28 and that another of $r_2 = 0.40$ has been obtained from a sample of size 23. Are they significantly different?

We find $X_1 = 0.69315$, $X_2 = 0.42365$

$$x = 0.26950 \div \left(\frac{1}{25} + \frac{1}{20}\right)^{1/2} = \frac{0.26950}{(0.09)^{1/2}} = 0.898$$

which is not significant.

EXERCISES

1. In the manufacture of a certain article it is known that 2 per cent of the articles are defective What is the probability that a random sample of 10 of the articles will contain (*a*) no defectives, (*b*) exactly 1 defective, (*c*) exactly 2 defectives, (*d*) not more than 2 defectives?

2. If the probability that a person 65 years old will die within 1 year is 0.04, find the probability that of 5 persons 65 years old exactly 1 will die during the year.

3. A certain disease has a fatality of 10 per cent The records of a hospital show that, out of 15 patients belonging to a designated occupational class and admitted with this disease, 4 died. Does this indicate a lack of resistance to the disease on the part of this occupational class?

4. The presidents of the United States, up to and including Franklin D. Roosevelt, have a total of 70 sons and 46 daughters (*World Almanac*). Is this an unusual proportion if the ratio of male births to total births in the population at large is 0 51?

5. If a baseball player has a batting average of 0 300, what is the probability that he will get at least 25 hits out of 100 times at bat?

6. In a certain university the number of failures in freshman English, over a period of years, is 8 per cent In a given year there were 50 failures in a class of 500 Can this be attributed to chance?

7. Supposing that the observer who made the telescopic readings of exercise 1, page 24, is known to have a standard deviation in his readings of 0″ 52, and that the population mean is the same as that of the sample of 15 readings, find the probability of a deviation as great as or greater than the maximum deviation among the 15 readings, under the further assumption of a normal population *

8. Fit a normal curve to the data of Table B, page 8, finding the theoretical frequencies of the various classes.

9. (*a*) Fit a normal curve to the total frequencies of carapace length in Table P, page 63 (*b*) Fit a normal curve to the total frequencies of right chela length in this table.

10. (*a*) Fit a normal curve to the distribution of weights obtained as marginal totals in either Table R_1 or Table R_2 (*b*) Fit a normal curve to the distribution of heights obtained as marginal totals in either Table R_2 or Table R_3

11. Suppose that it has been determined that the average pulse rate of males in the 20–25 year age group is 72 beats per minute and that the standard deviation is 9.5 beats per minute. If a group of 25 distance runners, all in the given age group, were examined and found to have an average pulse rate of 65, should this be regarded as a significant deviation from the general average?

12. Suppose that the standard deviation of stature in men is 2 48 in. One hundred male students in a large university are measured and their average height is found to be 68 52 in. Determine the 98 per cent confidence limits for the mean height of the men of the university.

13. If 50 freshmen in a given university are found to have a mean height of 68 60 in , and 40 seniors a mean height of 69 51 in , is the evidence conclusive that the mean height of the seniors is greater than that of the freshmen? Assume the standard deviation of height to be 2 48 in

14. A certain intelligence test has been administered to a large group of pupils and it has been found that the standard deviation of scores is 38.5 for girls and 35.2 for boys. The test is given to a group of 17 girls, who make an average score of 185 1, and to a group of 23 boys, who make an average score of 156 7 Is there a significant difference in intelligence between these two groups?

15. One thousand articles from a factory are examined and found to be 3 per cent defective. Fifteen hundred similar articles from a second factory are found to be only 2 per cent defective Can it reasonably be concluded that the product of the first factory is inferior to that of the second?

16. In 1910, in the original registration states, the number of white males dying between the ages of 30 and 31 was 1609 out of a population of 253,445 of white males in this age group; the corresponding figures for white females of the same age were 1302 out of 239,912 ("United States Life Tables"). Is

* See Paul R. Rider, "Criteria for Rejection of Observations," Washington University Studies, new series, Science and Technology, No. 8, St Louis, 1933.

there a significant difference between the death rates of the two sexes at this age?

17. In 1910, in the original registration states, the number of negro males dying between the ages of 30 and 31 was 115 out of 6975. Can it be assumed that there is a racial difference in death rates at this age? (See preceding exercise for data on white males.)

18. In 1910, for white males in the age group 30–34 years, the number dying in Chicago was 902 out of 106,307; in New York 2130 out of 221,598 ("United States Life Tables.") Are the death rates in the two cities significantly different?

19. How significant are the results of vaccination shown in Table S?

TABLE S

SMALLPOX DATA, LONDON, 1901

(Macdonell, *Biometrika*, vol. 1)

	Recoveries	Deaths
Vaccinated	652	108
Unvaccinated	96	98

20. A correlation coefficient of 0.5 is discovered in a sample of 19 pairs. Is this significantly different (*a*) from 0.3, (*b*) from 0?

21. Answer the questions of the foregoing exercise for the case in which the coefficient is a partial correlation coefficient $r_{12.345}$.

22. The correlation coefficient between mathematics aptitude and language aptitude for a group of 20 boys is 0.42. For a group of 25 girls the correlation is 0.75. Is the difference significant?

23. How probable is it that the correlation in the population from which the sample shown in Table J, page 43, was taken is 0.50 or greater? (See exercise 1, page 62.)

24. Is there a significant sex difference in correlation between red blood cells and hemoglobin as determined from Table O, page 63? (See exercise 2, page 62.)

CHAPTER VI

STUDENT'S DISTRIBUTION

40. Student's distribution and the reliability of a mean. We have seen in the preceding chapter that, if we know the standard deviation σ of a normal population, we can test the probability that a sample mean $\bar{X}$ deviates by more than a specified amount from the population mean μ by treating the quantity $N^{1/2}(\bar{X} - \mu)/\sigma$ as a normal deviate with unit standard deviation. If, however, we do not know the standard deviation of the population, it becomes necessary to estimate it from the sample. We might use the standard deviation of the sample,* viz., $s = [\Sigma(X - \bar{X})^2/N]^{1/2}$, but for certain reasons it is more desirable to use †

$$s[N/(N - 1)]^{1/2} \tag{1}$$

Then the estimated standard deviation of the mean, found by dividing (1) by $N^{1/2}$, is $s/(N - 1)^{1/2}$, and if we set

$$t = \frac{\bar{X} - \mu}{s/(N - 1)^{1/2}} = \frac{\bar{X} - \mu}{[\Sigma(X - \bar{X})^2/N(N - 1)]^{1/2}} \tag{2}$$

* Any quantity such as a standard deviation, a median, a correlation coefficient, when calculated from a sample, is called a *statistic*; the corresponding quantity in the population is called a *parameter*.

† The expression (1) is the value of σ which will make the value of s obtained from the sample the most probable It is obtained by the method which Fisher terms that of *maximum likelihood* and is called the *optimum* value of σ (See section 50; also Deming and Birge, "On the statistical theory of errors," Graduate School of the U S. Department of Agriculture, Washington, reprinted with additional notes dated 1937 from *Reviews of Modern Physics*, vol 6, 1934, pp. 119–161)

then the quantity t is not distributed normally, but in "*Student's*" *distribution*

$$Ydt = \frac{[(n-1)/2]!}{(n\pi)^{1/2}[(n-2)/2]!}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} dt \tag{3}$$

with $n = N - 1$, the so-called number of *degrees of freedom.*

The symbol $k!$, called "k factorial," is defined as

$$\int_0^\infty x^k e^{-x} dx \tag{4}$$

This reduces to $k(k-1)(k-2) \ldots 3 \cdot 2 \cdot 1$ if k is an integer. The integral (4) is often called the gamma function, $\Gamma(k+1)$, and we have the relation $k! = \Gamma(k+1)$.

Any quantity t *which is the ratio of a normal deviate to a stochastically independent* * *estimate of its standard deviation, obtained from samples from a normal population, is distributed in Student's distribution with* n *equal to the number of degrees of freedom utilized in estimating the standard deviation.*†

The variance of t is $n/(n-2)$, and as n becomes larger the distribution approaches a normal distribution. For $n > 30$ it is permissible to use tables of the normal distribution by taking $t[(n-2)/n]^{1/2}$ as a normal deviate with unit standard deviation, and for $n > 100$ the quantity t itself may be considered as the deviate.

The probability that t will be numerically greater than a specified value has been tabulated,‡ but in practice one generally uses tables such as Table V at the end of this book, in which t is tabulated in terms of the probability that it will be exceeded in random samples.

In order to gain an idea of the use of Student's distribution let us suppose that a machine which produces mica insulating washers for use in electrical devices is set to turn out washers having a thickness of 10 mils (1 mil = 0.001 inch). A sample of 10 washers has an average thickness of 9.52 mils with a standard deviation

* I.e , independent in a probability sense

† See R. A. Fisher, "Applications of 'Student's' distribution," *Metron*, vol. 5, 1925, No. 3, pp 90–104.

‡ See *Metron*, vol. 5, 1925, No. 3, pp. 105–120.

of 0.60 mil. Let us see the significance of such a deviation. We find by using (2) that

$$t = \frac{9.52 - 10}{0.60/9^{\frac{1}{2}}} = -2.4, \quad n = 10 - 1 = 9$$

The probability that t is numerically greater than 2.4, or that the mean thickness of a sample of 10 washers will deviate more than 2.4 tımes its estımated standard deviation above or below the populatıon mean (here assumed to be 10 mils), is, from the tables ın *Metron* just referred to, found to be

$$P(|t| > 2.4) = 0.04$$

For such a probabılty we should be inclıned to say that the deviation is not altogether due to chance. In practice we should merely note ın Table V that $2.4 > 2\ 262$, the 5 per cent level for $n = 9$.

41. Confidence limits for the population mean.[*] The foregoing probabılıty statement can be phrased as follows: If 10 ıs the mean thickness, then the probabılity of obtaining a sample of 10 with a mean and a standard devıation leadıng to a more ımprobable value of t is 0.04. A sımılar probabılıty statement can be made on the basis of the assumption that the mean thickness of the population of washers ıs 9, or 11, or any other value.

Suppose that we wish to estımate, from our sample, lımıts within whıch the population mean lies, with some assurance that we are correct. For example, suppose that we want to be 98 per cent sure (in the sense that we should be correct in our judgment 98 per cent of the time). We can set $P(|t| \geqq t_1) = 0.02$, and for 9 degrees of freedom find, from Table V, that $t_1 = 2.821$. That ıs,

$$\frac{|9.52 - \mu|}{0.60/9^{\frac{1}{2}}} \geqq 2.821, \quad |9.52 - \mu| \geqq 0.5642$$

$$8.9558 \leqq \mu \leqq 10.0842$$

[*] For a clear elementary dıscussion of the subject of confidence lımits and fiducıal probabılıty the reader ıs referred to H L Rıetz, "On a recent advance ın statıstıcal ınference," *Amerıcan Mathematıcal Monthly*, vol 45, 1938, pp. 149–158.

If μ is less than 8.9558 or greater than 10.0842, then $|t|$ found from our sample values of $\bar{X}$ and s will exceed 2.821, and therefore, if any such value of μ is chosen as a hypothetical mean, the sample will be regarded as inconsistent at the 2 per cent level of significance, even though this value of μ may be the true one. On the other hand, if any value of μ for which $8.9558 \leqq \mu \leqq 10.0842$ is chosen as a hypothetical mean, then for this choice $|t|$ will be less than 2.821 and the sample will be regarded as consistent with such a hypothesis at the 2 per cent level of significance. But, whatever be the true value of μ, in repeated samples with μ set equal to this value, we shall obtain simultaneous values of $\bar{X}$ and s such that $|t| \leqq 2.821$ in 98 per cent of cases, so that whatever be the value of μ we expect consistent samples at the 2 per cent level of significance in 98 per cent of cases. Therefore, if we suppose our sample to be consistent (at the 2 per cent level of significance) with regard to the unknown value of μ from which it did arise, then in 98 per cent of cases in the long run we shall be correct in our supposition. Consequently, in the present instance we suppose our sample to be consistent, which means that we suppose $8.9558 \leqq \mu \leqq 10.0842$; and although we may be wrong in this instance, yet 98 per cent of the time that we write such inequalities, based on the values of X and s observed, we shall be right; for 98 per cent of the time we shall have drawn a consistent sample, whatever μ.

The values 8.9558 and 10.0842 may be termed 98 per cent *fiducial* or *confidence limits* of μ corresponding to the observed sample.

42. Testing the significance of the difference between two means. If we wish to determine the significance of the difference between two means of small samples we test the hypothesis that they came from the same normal population. If two variables having variances σ_1^2 and σ_2^2 respectively are uncorrelated, the variance of their difference is $\sigma_1^2 + \sigma_2^2$. Thus the variance of the difference between two means $\bar{X}_1$ and $\bar{X}_2$ of samples of N_1 and N_2 respectively, taken from a population whose variance is σ^2, is $\sigma^2/N_1 + \sigma^2/N_2 = \sigma^2(1/N_1 + 1/N_2)$. If we do not know the population variance we estimate it from the expression

$$s'^2 = \frac{\Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2}{N_1 + N_2 - 2} = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2} \qquad (5)$$

where s_1^2 and s_2^2 are the variances of X_1 and X_2 respectively. The denominator of the foregoing fraction is the number of degrees of freedom used in estimating σ^2. To obtain this we must deduct two from the number of observations, one degree having been used up in calculating $\overline{X}_1$, another in calculating $\overline{X}_2$. The estimate of the standard deviation of $\overline{X}_1 - \overline{X}_2$ would then be $s'(1/N_1 + 1/N_2)^{1/2}$ and we write

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s'\left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)^{1/2}} = \frac{\overline{X}_1 - \overline{X}_2}{\left(\dfrac{N_1 + N_2}{N_1 + N_2 - 2}\right)^{1/2}\left(\dfrac{s_1^2}{N_2} + \dfrac{s_2^2}{N_1}\right)^{1/2}} \qquad (6)$$

Then t is distributed in Student's distribution with the number of degrees of freedom $n = N_1 + N_2 - 2$. If $N_1 = N_2 = N$, (6) reduces to the simpler form

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s'\left(\dfrac{2}{N}\right)^{1/2}} = \frac{\overline{X}_1 - \overline{X}_2}{\left(\dfrac{s_1^2 + s_2^2}{N - 1}\right)^{1/2}} \qquad (7)$$

with $n = 2(N - 1)$.

As an illustration let us consider the following problem: The ash content of coal from two different mines was analyzed, five analyses being made of the coal from the first mine, four of that from the second mine. Are we justified in supposing that the two mines consist of coal with the same percentage ash content on the basis of the results obtained, which are recorded in Tables 16A and 16B?

$$N_1 = 5, \quad \overline{X}_1 = \frac{107.5}{5} = 21.5, \quad N_1 s_1^2 = 30.02$$

$$N_2 = 4, \quad \overline{X}_2 = \frac{72}{4} = 18, \quad N_2 s_2^2 = 7.78$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\left(\dfrac{N_1 + N_2}{N_1 + N_2 - 2}\right)^{1/2}\left(\dfrac{N_1 s_1^2 + N_2 s_2^2}{N_1 N_2}\right)^{1/2}}$$

$$= \frac{21.5 - 18.0}{\left(\frac{9}{7}\right)^{1/2}\left(\frac{30.02 + 7\ 78}{5 \times 4}\right)^{1/2}}$$

$$= \frac{3.5}{(2.43)^{1/2}} = 2.245+$$

$$n = N_1 + N_2 - 2 = 7$$

$$P(|t| > 2.245) = 0.06$$

TABLE 16A

COAL FROM FIRST MINE

Per cent ash content X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$
24 3	2 8	7 84
20 8	−0 7	0 49
23 7	2 2	4 84
21 3	−0 2	0 04
17 4	−4 1	16 81
107 5	0 0	30 02

TABLE 16B

COAL FROM SECOND MINE

Per cent ash content X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
18 2	0 2	0 04
16 9	−1 1	1 21
20 2	2 2	4 84
16 7	−1 3	1 69
72 0	0 0	7 78

Thus, there are about 6 chances in 100 of observing a greater difference in percentage ash content, on the supposition that the mines are equal in this respect, and at the 5 per cent level of significance we can not judge that any difference between the mines has been detected.

43. Testing the significance of a regression coefficient. Suppose that we have fitted a regression line $Y' = a + bX$ or $y' = bx$ to a set of N pairs of values of X and Y and wish to see whether the regression coefficient b differs significantly from some hypothetical value β. We have previously found that

$$b = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

which is a weighted sum of y's, the weight corresponding to y_i being $w_i = x_i/\Sigma x^2$. If now we assume that, for a given value of x, y is normally distributed with variance σ^2, b will be normally distributed with variance

$$\sigma_b^2 = \sigma^2 \Sigma w^2 = \frac{\sigma^2}{(\Sigma x^2)^2} \Sigma x^2 = \frac{\sigma^2}{\Sigma x^2} \tag{8}$$

If we do not know σ^2 we must estimate it from the data. As an estimate we use

$$s'^2 = \frac{\Sigma(Y - Y')^2}{N - 2} = \frac{1}{N - 2}(\Sigma Y^2 - a\Sigma Y - b\Sigma XY) \tag{9}$$

the divisor $N - 2$ being the number of degrees of freedom left after the two constants a and b have been calculated.

If we divide $b - \beta$ by the estimate of the standard deviation of b we have a quantity distributed in Student's distribution with $N - 2$ degrees of freedom. The estimated value of σ_b is

$$\frac{s'}{(\Sigma x^2)^{1/2}} = \left[\frac{\Sigma(Y - Y')^2}{(N - 2)\Sigma(X - \bar{X})^2}\right]^{1/2} \tag{10}$$

and we set

$$t = (b - \beta) \div \left[\frac{\Sigma(Y - Y')^2}{(N - 2)\Sigma(X - \bar{X})^2}\right]^{1/2}, \quad n = N - 2 \tag{11}$$

For the illustration used in section 18 we have

$$b = 0.376, \quad \Sigma(Y - Y')^2 = 3.6062,$$

$$\Sigma(X - \bar{X})^2 = 110 - \tfrac{324}{5}, \quad N - 2 = 3$$

To test whether b is significantly different from zero we set $\beta = 0$ in (11) and find

$$t = 0\ 376 \div \left[\frac{3\ 6062}{3(110 - 324/5)}\right]^{1/2} = \frac{0.376}{0.1631} = 2.31$$

This is not significant, as the 5 per cent value of t for 3 degrees of freedom is 3.182.

44. Testing the significance of the difference between two regression coefficients. Suppose that we have two regression equations

$$Y'_1 = a_1 + b_1X, \quad Y'_2 = a_2 + b_2X \tag{12}$$

calculated from N_1 and N_2 pairs of values respectively. To test the significance of the difference between b_1 and b_2 we calculate

$$s'^2 = \frac{\Sigma(Y_1 - Y'_1)^2 + \Sigma(Y_2 - Y'_2)^2}{N_1 + N_2 - 4} \tag{13}$$

the denominator being the total number of degrees of freedom, $(N_1 - 2) + (N_2 - 2)$. The estimated variances of b_1 and b_2 are respectively

$$\frac{s'^2}{\Sigma x_1^2} = \frac{s'^2}{\Sigma X_1^2 - (\Sigma X_1)^2/N_1} \tag{14}$$

$$\frac{s'^2}{\Sigma x_2^2} = \frac{s'^2}{\Sigma X_2^2 - (\Sigma X_2)^2/N_2} \tag{15}$$

and the estimated variance of the difference $b_1 - b_2$, under the assumption of no correlation between b_1 and b_2, is the sum of the foregoing. We test the difference $b_1 - b_2$ by setting

$$t = \frac{b_1 - b_2}{s'\left(\frac{1}{\Sigma x_1^2} + \frac{1}{\Sigma x_2^2}\right)^{1/2}}, \quad n = N_1 + N_2 - 4 \tag{16}$$

and continuing as in the preceding sections.

45. Testing the significance of a partial regression coefficient. Suppose that we have fitted a multiple regression equation

$$Y' = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{17}$$

and wish to test the significance of one of the regression coefficients, say b_i. It can be shown, by the method used in the preceding section, that the estimated variance of b_i is s'^2c_{ii}, where

$$s'^2 = \frac{\Sigma(Y - Y')^2}{N - k - 1} \tag{18}$$

and c_{ii} is found as in section 20 of Chapter III. The denominator in (18) is the number of sets of values used in obtaining the regression equation diminished by the number of constants in the equation. It is the number of degrees of freedom used in the estimate

s'^2. Then to test the significance of the deviation of b_i from a hypothetical value β_i (e.g., zero) we set

$$t = \frac{b_i - \beta_i}{s'c_{ii}^{\frac{1}{2}}} = (b_i - \beta_i) \div \left[\frac{c_{ii}\Sigma(Y - Y')^2}{N - k - 1}\right]^{\frac{1}{2}} \tag{19}$$

and use probability tables of the quantity t with $n = N - k - 1$.

In the example of multiple regression worked out in Chapter III we found

$$\begin{array}{lll} c_{00} = 27.9031 & c_{01} = -\ 3.1290 & c_{02} = -\ 6.3226 \\ & c_{11} = 0.3613 & c_{12} = 0.7032 \\ & & c_{22} = 1.4581 \\ b_0 = -\ 5.9357 & b_1 = 1.2194 & b_2 = 1.7484 \end{array}$$

We need

$$\begin{aligned} \Sigma(Y - Y')^2 &= \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma X_1Y - b_2\Sigma X_2Y \\ &= 55 + 5.9357 \times 15 - 1.2194 \times 71 - 1.7484 \times 32 \\ &= 1.5093 \end{aligned}$$

To test whether b_1 has a significant value, that is, whether it is significantly different from zero, we set

$$t = 1\,2194 \div \left(\frac{0.3613 \times 1.5093}{5 - 3}\right)^{\frac{1}{2}} = \frac{1.2194}{1.6512} = 0.739$$

From Table V we find for $n = 2$ that the probability that t will exceed this value numerically is between 0.5 and 0.6, so the value of b_1 is not at all significant—we should expect a value this large numerically more than half the time.

46. Comparing two partial regression coefficients. To compare two partial regression coefficients in the same regression equation,* such as b_1 and b_2, we set t equal to their difference divided by the square root of

$$s'^2(c_{11} - 2c_{12} + c_{22}) \tag{20}$$

* See Fisher, "Statistical Methods for Research Workers," section 29.

In the present example we have

$$t = \frac{b_2 - b_1}{s'(c_{11} - 2c_{12} + c_{22})^{1/2}}$$

$$= (1.7484-1.2194) \div \left[\frac{1\ 5093}{2}(0.3613-2\times 0.7032+1.4581)\right]^{1/2}$$

$$= \frac{0.5290}{(0.311670)^{1/2}} = \frac{0.5290}{0.5583} = 0.947, \quad n = 2$$

and Table V shows that there is no significant difference, the probability of a greater difference being near 0.5.

47. Testing whether a sample has been drawn from uncorrelated material. If we are sampling from uncorrelated material the values of r obtained follow the rather simple distribution

$$Ydr = \frac{[(N - 3)/2]!}{\pi^{1/2}[(N - 4)/2]!}(1 - r^2)^{(N-4)/2}\,dr \qquad (21)$$

where N is the number of pairs of values in the sample. By means of the substitution

$$r = \frac{t/n^{1/2}}{(1 + t^2/n)^{1/2}}, \quad dr = \frac{dt}{n^{1/2}(1 + t^2/n)^{3/2}}, \quad n = N - 2 \quad (22)$$

the distribution (21) is transformed into Student's distribution (3) of section 40 with $n = N - 2$.

To test whether a value of r obtained from a sample of N pairs of values is significantly different from zero, that is, whether the material from which we are sampling is correlated, we make the inverse of transformation (22), viz.,

$$t = \frac{n^{1/2}r}{(1 - r^2)^{1/2}}, \quad n = N - 2 \qquad (23)$$

and use tables of Student's distribution.

In section 22 of Chapter IV we found $r = 0.80$ from 5 pairs of values of X and Y. Using formula (23) we find

$$t = \frac{3^{1/2} \times 0.80}{(1 - 0.64)^{1/2}} = \frac{1.3856}{0.6} = 2.310, \quad n = 3$$

The probability that t is numerically greater than 2.310, which is the probability that r deviates from zero by more than ±0.80, is about 0.1. The observed value of r therefore could not be regarded as significant The reason why such a comparatively high value of the correlation coefficient is not significant is that it is obtained from such a small number of cases, that is, the number of degrees of freedom is small.

48. Testing the significance of a partial correlation coefficient. If we have found a partial correlation coefficient $r_{12\,34\quad k}$ from N sets of values, the significance of its difference from zero can be obtained by using the transformation employed in the preceding section, with $n = N - k$. For example, we found in section 29 a value $r_{12\,3} = -\ 0.977$ from 5 sets of values. In this case $N = 5$, $k = 3$, $n = N - k = 2$, and

$$t = \frac{2^{1/2}(-0\ 977)}{[1 - (-0.977)^2]^{1/2}} = -\ 6\ 479, \quad n = 2$$

The probability that t will be larger than this numerically is slightly greater than 0.02 and the corresponding value of r may be regarded as significant, although not highly so.

EXERCISES

1. A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressure. 5, 2, 8, −1, 3, 0, 6, −2, 1, 5, 0, 4. Can it be concluded that the stimulus will be in general accompanied by an increase in blood pressure?

2. Using the observations of exercise 1, page 24, set 95 per cent fiducial limits for the vertical angular diameter of Venus

3. Below are given the gains in weight (pounds) of hogs fed on two different diets Twelve animals were fed on diet A, 15 on diet B. Is either diet superior?

Gains in weight on diet A: 25, 30, 28, 34, 24, 25, 13, 32, 24, 30, 31, 35

Gains in weight on diet B: 44, 34, 22, 8, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

4. Test for significance the regression coefficient found in exercise 2, page 43. Can the population regression coefficient be regarded as great as 0 5?

5. Test for significance the partial regression coefficients found in exercise 8, page 45

6. A correlation coefficient of 0.42 is found in a sample of 25 pairs of values. Can it be regarded as significantly different from zero?

7. If the coefficient in the preceding exercise is a partial correlation coefficient of order 3 obtained from 25 sets of 5 values each, can it be regarded as significant?

8. (*a*) Is there a significant sex difference in mean number of blood cells as determined from Table O, page 63? (*b*) Is there a significant sex difference in mean amount of hemoglobin as determined from this table?

9. Find the coefficients of regression of hemoglobin on red blood cells for the men and for the women of Table O, page 63. Is there a significant difference between these coefficients?

CHAPTER VII

THE CHI-SQUARE DISTRIBUTION

49. Chi square. Suppose that the variable X has the distribution $(2\pi)^{-\frac{1}{2}} \exp(-X^2/2\sigma^2)dX/\sigma$, in other words that it is normally distributed about zero with variance σ^2. If we have n values of X, we define

$$\chi^2 = \frac{\sum_{i=1}^{n} X_i^2}{\sigma^2} \tag{1}$$

In words, chi square is the sum of squares of n independent normal deviates divided by their common variance. The number of independent deviates, n, is called the number of *degrees of freedom.*

It can be shown that χ^2 has the distribution

$$\frac{1}{2^{n/2}[(n-2)/2]!} e^{-\chi^2/2} (\chi^2)^{(n-2)/2} d(\chi^2) \tag{2}$$

and tables of integrals of this function, which is a Pearson type III function, have been prepared. They show the probabilities that χ^2 will exceed certain values,* or the values of χ^2 that will be exceeded certain fixed proportions of times.† Table VI at the end of the book is of the latter type.

For large values of n, say $n > 30$, χ (not χ^2) may be regarded as being normally distributed, with mean $(n - \frac{1}{2})^{\frac{1}{2}}$ and standard deviation $2^{-\frac{1}{2}}$; that is, the quantity $(2\chi^2)^{\frac{1}{2}} - (2n - 1)^{\frac{1}{2}}$ may be used as a normal deviate with unit standard deviation.

A very important property of the chi-square distribution is that *the sum of any number of independent quantities each of which is*

* Table XII in Karl Pearson's "Tables for Statisticians and Biometricians," part 1.

† Table III in R. A. Fisher's "Statistical Methods for Research Workers."

distributed in a chi-square distribution is itself distributed in a chi-square distribution with degrees of freedom equal to the sum of the degrees of freedom of the separate components.

50. Distribution of variances and standard deviations. If in formula (2) of section 49 we set

$$\chi^2 = \frac{Ns^2}{\sigma^2}, \quad d(\chi^2) = \frac{N}{\sigma^2}\, d(s^2), \quad n = N - 1 \tag{3}$$

where s^2 is the variance of a sample of N individuals, the chi-square distribution is transformed into

$$\frac{N^{(N-1)/2}}{2^{(N-1)/2}[(N-3)/2]!\,\sigma^{N-1}}\,(s^2)^{(N-3)/2}e^{-Ns^2/2\sigma^2}d(s^2) \tag{4}$$

which is the distribution of variances of samples of size N from normal material. This can readily be changed into the distribution of standard deviations if desired.

The above connection between chi square and variance can be used in comparing a sample variance with the population variance. Suppose, for example, that a sample of 9 individuals has a variance of 4.8 How likely is it that the sample came from a population having a variance of 3.2?

We set

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{9 \times 4.8}{3.2} = 13.5, \quad n = N - 1 = 8$$

From Table VI, for 8 degrees of freedom, we find

$$P(\chi^2 > 13.362) = 0.10$$

Thus, a population having a variance of 3 2 would yield a sample having a variance of 4 8 or greater about once out of 10 times.

From equation (4) can be obtained the value of σ^2 which will make the observed value of s^2 the most probable. This value is called the *optimum* value of σ^2, and the method of obtaining it is called the method of *maximum likelihood.* The method consists of differentiating (4) with respect to σ^2, setting the derivative equal to zero, and solving the resulting equation. This gives

$$\sigma^2 = \frac{N}{N-1}\, s^2 \tag{5}$$

For this reason the quantity

$$s'^2 = \frac{N}{N-1}\, s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{6}$$

is regarded as a better estimate of the population variance than is the sample variance s^2 itself. Also (6) is the mean value of the variance of samples of size N.

51. Testing the homogeneity of several estimated variances.* The chi-square distribution can be used as an approximate test of the homogeneity of several estimates of variance. Suppose that we have k independent estimates of variance,

$$s_1'^2 = \frac{1}{n_1}\,\Sigma(X_1 - \bar{X}_1)^2, \quad s_2'^2 = \frac{1}{n_2}\,\Sigma(X_2 - \bar{X}_2)^2, \cdots,$$
$$s_k'^2 = \frac{1}{n_k}\,\Sigma(X_k - \bar{X}_k)^2 \tag{7}$$

based upon $n_1, n_2, \ldots, n_k$ degrees of freedom respectively. The pooled variance is the weighted mean

$$s'^2 = \frac{1}{n}\,\Sigma n_i s_i'^2, \quad n = \Sigma n_i \tag{8}$$

Then the quantity

$$\frac{1}{C}\,(n \log_e s'^2 - \Sigma n_i \log_e s_i'^2) = \frac{2.3026}{C}\,(n \log_{10} s'^2 - \Sigma n_i \log_{10} s_i'^2) \tag{9}$$

in which

$$C = 1 + \frac{1}{3(k-1)}\left(\Sigma \frac{1}{n_i} - \frac{1}{n}\right) \tag{10}$$

is approximately distributed as χ^2 with $k - 1$ degrees of freedom, and an unduly large value of χ^2 will indicate the presence of discrepancies among the several estimates of variance (The special case in which only two estimates are involved is treated in the next chapter.) Since $C > 1$, it need not be calculated if χ^2 for $C = 1$ is *non-significant.*

For example, if we have three estimates of variance, 4.2, 6.0,

* See M S Bartlett, "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society of London,* series A, vol. 160, 1937, pp. 273 ff.

and 3.1, based on 4, 5, and 11 degrees of freedom respectively, we can form a table such as Table 17, from which we find

TABLE 17

i	$s_i'^2$	n_i	$n_i s_i'^2$	$\log_e s_i'^2$	$n_i \log_e s_i'^2$
1	4 2	4	16 8	1 43508	5 74032
2	6 0	5	30 0	1 79176	8 95880
3	3 1	11	34 1	1 13140	12 44540
Total		20	80 9	4 35824	27 14452

$$s'^2 = \frac{\Sigma n_i s_i'^2}{\Sigma n_i} = \frac{80.9}{20} = 4.045$$

$$n \log_e s'^2 = 20 \times 1.39748 = 27.94960$$

$$C = 1 + \frac{1}{3(3-1)}\left(\frac{1}{4} + \frac{1}{5} + \frac{1}{11} - \frac{1}{20}\right) = 1.0818$$

$$\chi^2 = \frac{1}{C}\,(n \log_e s'^2 - \Sigma n_i \log_e s_i'^2)$$

$$= \frac{27.94960 - 27.14452}{1\ 0818} = \frac{0\ 80508}{1.0818} = 0.744$$

The probability of a greater value than this, for 2 degrees of freedom, is between 0.50 and 0.70, so that the three estimates of variance may be regarded as homogeneous. Actually the use of C is unnecessary in this example.

52. Small samples from binomial and Poisson distributions. Quite similar to the use of the chi-square test described in section 50 is its use in testing the homogeneity of small samples of material supposed to have come from a binomial or from a Poisson exponential distribution, by comparing the variance of the sample with the expected population variance as estimated from the sample.

A single member of the binomial distribution $(q + p)^N$ will be understood to mean the number of occurrences in N independent trials or observations, in each of which the probability of occurrence is p. A sample of k members is of the form $X_1, X_2, \ldots, X_k$, in which each X is an integer between 0 and N inclusive. Then the *index of dispersion*

$$\frac{\Sigma(X - \overline{X})^2}{Np'q'} = \frac{\Sigma(X - \overline{X})^2}{N\frac{\overline{X}}{N}\left(1 - \frac{\overline{X}}{N}\right)} \tag{11}$$

is distributed approximately as χ^2 with $k - 1$ degrees of freedom. It may be noted that the denominator of (11) is the expected value of the population variance as estimated from the sample, and that the quantity (11) is comparable to the χ^2 in (3) of section 50.

Suppose, by way of illustration, that a plant pathologist is investigating the distribution of a certain plant disease and has divided a field into plots, in each of which is a certain number of plants. If every plant in the field has an equal and independent chance of becoming infected, we expect that the index of dispersion will not have an unusual value. If, on the other hand, there is a deviation from a random distribution of infection, as for instance if the plots on one side of the field show a higher degree of infection, we may expect to get an unusual value for the index.

In an experiment of the type just described, a field was divided into 12 plots, each of which contained 90 plants. The numbers of infected plants from the various plots were as follows: 19, 6, 9, 18, 15, 13, 14, 15, 16, 20, 22, 14. Is the infection random? That is, does the variability conform to expectation?

Here we have

$$\overline{X} = \frac{\Sigma X}{k} = \frac{181}{12} = 15.083$$

$$\Sigma(X - \overline{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{k} = 2953 - \frac{(181)^2}{12} = 222.9167$$

The index of dispersion, as calculated from (11), in which $N = 90$, is

$$\chi^2 = 17.76$$

The number of degrees of freedom is 11, one less than the number of plots. Table VI shows that the probability of obtaining this value of χ^2, or a larger value, is between 0 05 and 0.10, which would indicate that groups of infected plants do not occur together oftener than would happen by chance.

The index of dispersion for the Poisson exponential distribution is

$$\frac{\Sigma(X - \bar{X})^2}{\bar{X}} \tag{12}$$

The denominator is the expected value of the population variance as estimated from the sample, since in the Poisson distribution the variance is equal to the mean.

This index is useful in checking whether the variability in bacterial counts made by the dilution method is that which is to be expected. According to theory, if the technique of dilution provides a random distribution of organisms, and if these can develop on the plate independently, then the numbers of colonies counted on plates from the same dilution are distributed in a Poisson exponential distribution. An unusual value of χ^2 obtained from (12) may be taken as an indication that the technique is not good.

Suppose that the following counts were obtained on 15 plates made with the same dilution of a bacterial culture: 193, 168, 161, 153, 183, 152, 171, 156, 159, 140, 151, 152, 133, 164, 157. Is the variability of the counts consistent with that to be expected according to the Poisson distribution?

We calculate the index of dispersion

$$\chi^2 = \frac{\Sigma(X - \bar{X})^2}{\bar{X}} = \frac{3229.73}{159.53} = 20.25-$$

For 14 degrees of freedom (that is, one less than the number of counts), this is not an unusual value, since Table VI shows that the probability of a greater value is between 0.1 and 0 2. We conclude that the variability is consistent with expectation.

53. Combining homogeneous estimates of correlation. If correlation coefficients have been calculated from two or more samples extracted from equally correlated populations, then their values can be combined to give a better estimate of the population correlation.

Suppose, for example, that we have found the correlations $r_1, r_2, \ldots, r_k$ in k independent samples of size $N_1, N_2, \ldots, N_k$, respectively. We make the transformations *

$$X_i = \frac{1}{2}\log_e\frac{1+r_i}{1-r_i}, \quad i = 1, 2, \ldots, k \tag{13}$$

The several X_i must be weighted inversely according to their variances. Since the variance of X_i is * $1/(N_i - 3)$, the weighted average of the X's is

$$X = \frac{\Sigma(N_i - 3)X_i}{\Sigma(N_i - 3)} \tag{14}$$

This may be changed back to a correlation coefficient by the transformation

$$X = \frac{1}{2}\log_e\frac{1+r}{1-r}, \quad \text{or } r = \frac{e^{2X}-1}{e^{2X}+1} = \tanh X \tag{15}$$

Before combining different values of r in this way, however, one should test the validity of the underlying assumption that the samples from which they were calculated came from equally correlated populations.

The correlation coefficient r_i is an estimate of a population correlation ρ_i appropriate to the ith sample, and if we make the transformations (13) then each X_i is approximately normally distributed with mean and variance

$$\overline{X}_i = \mu_i = \frac{1}{2}\log_e\frac{1+\rho_i}{1-\rho_i}, \quad \sigma^2_{X_i} = \frac{1}{N_i - 3} \tag{16}$$

respectively.

The hypothesis to be tested is that

$$\rho_i = \rho, \quad i = 1, 2, \ldots, k \tag{17}$$

in which ρ is a constant, or that

$$\overline{X}_i = \mu_i = \mu = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} \tag{18}$$

(It should be remarked here that the case in which $k = 2$ has already been treated in section 39.)

* See section 38

On this hypothesis we have k independent quantities, $X_i(i = 1,2, \ldots, k)$, distributed in an approximately normal manner with common mean μ and with variances given by the last part of (16). In a somewhat more general case the variance of X_i might be given by $\sigma^2/(N_i - 3)$, so that we are led to the general problem of finding an estimate of the variance σ^2 from a set of numbers which are normally distributed about a common mean, and whose variances are known fractions of that variance.

It is found * that s'^2, the estimate of σ^2, may be obtained from the equation

$$\begin{aligned}(k-1)s'^2 &= \Sigma(N_i - 3)(X_i - X)^2 \\ &= \Sigma(N_i - 3)X_i^2 - \frac{[\Sigma(N_i - 3)X_i]^2}{\Sigma(N_i - 3)} \qquad (19)\end{aligned}$$

in which X is the weighted mean of the X_i as given by (14).

Then the quantity

$$\chi^2 = \frac{(k-1)s'^2}{\sigma^2} \qquad (20)$$

is distributed as χ^2 with $k - 1$ degrees of freedom. Since $\sigma^2 = 1$ in the present case, our test reduces to testing the significance of the quantity (19), calling it χ^2 because of (20). If this χ^2 is significant, we judge that the variation among the X's is greater than is to be expected at the level of significance chosen, on the hypothesis that $\mu_i = \mu(i = 1, 2, \ldots, k)$. In such a situation we reject the hypothesis, which implies the rejection of the hypothesis $\rho_i = \rho$.

If χ^2 is not significant, we judge that the r_i are homogeneous and proceed to combine them as explained in the early part of this section. It will be noted that the quantities needed in forming the combined estimate have already been calculated in making the homogeneity test.

As an illustration, suppose that we have obtained a correlation coefficient of 0.60 from 33 pairs of values, another of 0 52 from 40

* See F Yates, "The analysis of multiple classifications with unequal numbers in the different classes," *Journal of the American Statistical Association*, vol 29, 1934, p 56.

pairs of values, and a third of 0.44 from 28 pairs of values. The calculations necessary for making the homogeneity test are shown in Table 18.

TABLE 18

i	r_i	X_i	$N_i - 3$	$(N_i - 3)X_i$	$(N_i - 3)X_i^2$
1	0 60	0 69315	30	20 79450	14 41371
2	0 52	0 57634	37	21 32458	12 29021
3	0 44	0 47223	25	11 80575	5 57503
Total			92	53 92483	32 27895

From (19) we find

$$\chi^2 = 32.27895 - \frac{(53.92483)^2}{92}$$

$$= 32.27895 - 31.60747 = 0.671$$

For two degrees of freedom this is not an unusual value, so that we conclude that the three sample values of the correlation coefficient may be regarded as coming from equally correlated populations.

To combine them we make use of (14), finding

$$X = \frac{53.92483}{92} = 0.58614$$

From the second part of (15) we get $r = 0.53$.

54. Test of goodness of fit. One of the principal uses of the chi-square distribution is testing how well an observed frequency distribution fits a theoretical distribution, although it affords only an approximate test in this instance.

If the observed frequency in a class of the distribution is f_0 and the theoretical or expected frequency is f, then

$$\chi^2 = \Sigma \frac{(f_0 - f)^2}{f} \tag{21}$$

The number of degrees of freedom is the number of classes less the number of constants in which we have forced the theoretical distribution to agree with the observed. For example, if we have made the totals agree and have caused the theoretical distribution to have the same mean and the same standard deviation as the observed, the number of degrees of freedom is three less than the number of classes. If the resulting value of χ^2 is unusual, say the probability is 0.01 or less, we conclude that, if the theoretical distribution which we have chosen represents the population adequately, then the data which we have observed are unusual—we should get a worse fit only once out of a hundred times as a matter of chance. In such a situation we are inclined to discard the hypothesis that the theoretical distribution is adequate. If we obtain an exceptionally small value of χ^2, say one such that the probability of getting a larger value is 0.99, the fit is too good and we suspect that the data are not random.

As an illustration of the method we shall consider the distribution of heights and the normal distribution with which it was fitted. (See section 32.) The details of the work are shown in Table 19.

TABLE 19

COMPUTATION OF χ^2

Theoretical frequency f	Observed frequency f_0	$f_0 - f$	$(f_0 - f)^2$	$\dfrac{(f_0 - f)^2}{f}$
0 1 1 2 12 0	1 2 9	− 1 3	1 69	0 1271
55 1	48	− 7 1	50 41	0 9149
114 2	131	16 8	282 24	2 4715
108 4	102	− 6 4	40 96	0 3779
45 4	40	− 5 4	29 16	0 6423
8 7 0 8	13	3 5	12 25	1 2895
345 9	346	0 1	416 75	5 8232

$\chi^2 = 5\ 8232, \quad n = 3$

$P(\chi^2 > 5.8232)$ is between 0 10 and 0.20 (Table VI)

It will be noticed that the first three classes have been combined into a single class. This is because the theoretical frequencies in the first two are small, it being an empirical rule that we should never use alone a class having in it a theoretical frequency less than five. There are six classes, but it will be recalled that we made the normal distribution agree with the observed distribution in total, mean, and standard deviation. Thus, the number of degrees of freedom is $6 - 3 = 3$. It is seen from Table VI that the probability of a greater value of χ^2 than that observed, viz , 5 8232, is between 0.10 and 0 20. The observed value can hardly be regarded as significant, since a larger value would occur by chance oftener than once in ten times.

55. Application to contingency tables. If a set of individuals is classified with respect to two or more different attributes and the frequencies tabulated we have a *contingency table.* As a simple example of a contingency table let us consider a 2 × 2 table (Table 20).*

TABLE 20

HAIR-COLOR AND EYE-COLOR (Observed)

Eye-color	Hair-color		Total
	Light	Dark	
Blue .	26	9	35
Brown..	7	18	25
Total .	33	27	60

Our problem is to determine whether there is any connection between hair-color and eye-color or whether they are independent of each other. In the case of a 2 × 2 table such as the one at hand this problem is perhaps best considered from the standpoint of the difference between two proportions. (See section 37.) However, we shall show how the χ^2 distribution can be used in testing this independence, as the method can be used for tables with a greater number of compartments.

* Cf. p. 82.

If the two attributes were quite independent we should expect the 35 blue-eyed persons to be distributed in the same proportion as the entire group, that is, we should expect $^{33}/_{60} \times 35(=19.25)$ of them to be in the light-haired class and $^{27}/_{60} \times 35(=15.75)$ of them to be in the dark-haired group Similarly we should expect the 25 brown-eyed persons to be divided into $^{33}/_{60} \times 25(=13.75)$ light-haired and $^{27}/_{60} \times 25(=11\ 25)$ dark-haired.* Thus we can form a table of values expected on the assumption that the two attributes are independent (Table 21).

TABLE 21

HAIR-COLOR AND EYE-COLOR (Expected)

Eye-color	Hair-color		Total
	Light	Dark	
Blue	19 25	15 75	35
Brown	13 75	11 25	25
Total.	33	27	60

From the table of observed and theoretical frequencies we can calculate

$$\chi^2 = \Sigma \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}}$$

$$= \frac{(26-19\ 25)^2}{19.25} + \frac{(9-15\ 75)^2}{15.75} + \frac{(7-13.75)^2}{13.75} + \frac{(18-11.25)^2}{11.25}$$

$$= \frac{(6.75)^2}{19.25} + \frac{(-6.75)^2}{15.75} + \frac{(-6.75)^2}{13.75} + \frac{(6.75)^2}{11.25} = 12.6234$$

The number of degrees of freedom is the number of compartments of the table which we are free to fill. In this case this is merely one,

* It should be noted that testing whether the proportions of light-haired and dark-haired are the same in each group (that is, blue-eyed and brown-eyed) is equivalent to testing whether the two groups are samples from the same population.

because if we fill one of them the others are determined by the marginal totals. In general, for a contingency table with r rows and c columns, the number of degrees of freedom is $(r - 1)(c - 1)$. For $n = 1$ we find

$$P(\chi^2 > 12.6234) < 0.01$$

so that in our example the departure from independence is decidedly significant.

If $n = 1$ the χ^2 distribution (2) reduces, after division by 2,* to

$$(2\pi)^{-\frac{1}{2}}e^{-\chi^2/2}d\chi$$

which is the normal distribution for χ. The probability that χ^2 is greater than k is

$$P(\chi^2 > k) = P(|\chi| > k^{\frac{1}{2}}) = 2P(\chi > k^{\frac{1}{2}}) = 2[1 - \varphi_{-1}(k^{\frac{1}{2}})] \quad (22)$$

It must be remembered that this applies only if n $= 1$.

In our example we find $(12.6234)^{\frac{1}{2}} = 3.55+$, and

$$P(\chi^2 > 12\,6234) = 2[1 - \varphi_{-1}(3.55)] = 2[1 - 0.9981] = 0.00038$$

It should be noted that the χ^2 test applied to a 2×2 contingency table always gives the same result as the method of testing the difference between two proportions given in section 37.

56. Contingency tables with small frequencies.† If the number in one or more compartments of the table is small, say less than 5, certain refinements of the above method yield better results. For example, consider Table 22, which shows the results of exposure of 20 people to a certain disease, 7 of the people having been inoculated and 13 not. The question to be answered is whether the inoculation is effective, that is, whether the frequencies in the various compartments differ significantly, on the whole, from the values that would be expected if they were distributed in pro-

* It is necessary to divide by 2, otherwise the area under the normal curve from $\chi = 0$ to $\chi = \infty$ (corresponding to $\chi^2 = 0$ and $\chi^2 = \infty$, respectively) would be 1 instead of $\frac{1}{2}$

† For a good discussion of this topic see F. Yates, "Contingency tables involving small numbers and the χ^2 test," *Supplement to the Journal of the Royal Statistical Society*, vol. 1 (1934), pp. 217–235; and J. O Irwin, "Tests of significance for differences between percentages based on small numbers," *Metron*, vol. 12, No. 2 (1935), pp. 1–94; also R. A. Fisher, "Statistical Methods for Research Workers", section 21.02.

portion to the marginal totals. In such a table we use Yates's correction, which consısts of addıng ½ to the smallest frequency of the table and adjusting the others so that the marginal totals will remain the same. Thıs ıs quite comparable to our procedure in approximatıng the point binomial by the normal curve, when, in order to find the probability of more than 10 occurrences, we use $1 - \varphi_{-1}[(10.5 - \mu)/(Npq)^{1/2}]$ rather than $1 - \varphi_{-1}[(10 - \mu)/(Npq)^{1/2}]$. The numbers 10, 3, 2, 5 in Table 22 would then be replaced by 9.5, 3.5, 2.5, 4.5, respectively, and the expected values would be

TABLE 22

RESULTS OF EXPOSURE TO A DISEASE

	Attacked	Not attacked	Total
Not ınoculated	10	3	13
Inoculated	2	5	7
Total	12	8	20

$$\tfrac{12}{20} \times 13 = 7.8, \quad \tfrac{8}{20} \times 13 = 5.2, \quad \tfrac{12}{20} \times 7 = 4.2, \quad \tfrac{8}{20} \times 7 = 2.8$$

As before,

$$\chi^2 = \frac{(9.5-7.8)^2}{7.8} + \frac{(3.5-5.2)^2}{5.2} + \frac{(2.5-4.2)^2}{4.2} + \frac{(4.5-2\ 8)^2}{2.8} = 2.64652$$

$$\chi = 1.627$$

$$P(\chi^2 > 2.64652) = 2P(\chi > 1.627)$$

$$= 2[1-\varphi_{-1}(1.627)] = 2(1-0.94813) = 0.10374$$

not a significant value.

Without Yates's correction we should have found

$$\chi^2 = \frac{(10 - 7.8)^2}{7.8} + \frac{(3 - 5.2)^2}{5.2} + \frac{(2 - 4.2)^2}{4.2} + \frac{(5 - 2.8)^2}{2.8} = 4.4322$$

$$\chi = 2.105+$$

$$P(\chi^2 > 4.4322) = 2P(\chi > 2.105)$$

$$= 2[1 - \varphi_{-1}(2.105)] = 2(1 - 0.98236) = 0.03528$$

It seems appropriate to consider at this point the exact treatment of a 2 × 2 table, which in general may be exhibited as Table 23.

TABLE 23

	A	Not A	Total
B	a	b	$a + b$
Not B	c	d	$c + d$
Total	$a + c$	$b + d$	N

It can be shown that when the marginal totals are fixed the probability of the observed values a, b, c, d in a contingency table is

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{(a+b+c+d)!\,a!b!c!d!} \tag{22}$$

If then we wish to find, in the above example (Table 22), the probability of the observed set of frequencies, or a set more extreme, viz.,

10	3	13
2	5	7
12	8	20

11	2	13
1	6	7
12	8	20

12	1	13
0	7	7
12	8	20

we calculate the sum

$$\frac{13!7!12!8!}{20!10!3!2!5!} + \frac{13!7!12!8!}{20!11!2!1!6!} + \frac{13!7!12!8!}{20!12!1!0!7!}$$

$$= \frac{13!7!12!8!}{20!}\left(\frac{1}{10!3!2!5!} + \frac{1}{11!2!1!6!} + \frac{1}{12!1!0!7!}\right)$$

$$= \tfrac{1}{9690}\,(462 + 42 + 1) = 0.0521156$$

This exact probability is comparable with $P(\chi > 1.627) = \frac{1}{2} \times 0.10374 = 0.05187$, using Yates's correction, and with $P(\chi > 2.105) = \frac{1}{2} \times 0.03528 = 0.01764$. It is thus seen that, in this particular case at least, Yates's correction yields much better results; in fact, without it our calculated probability is far from correct.

In conclusion we may say that a 2 × 2 contingency table can be dealt with by the method of the difference of two proportions or by the χ^2 method, but if the latter is used and any frequency is small, say less than 5, Yates's correction should be made. Better still is to employ the exact method.

For a fuller discussion of testing independence in contingency tables the reader should consult the references given earlier in the section.

EXERCISES

1. An entomologist sprayed 10 batches of 100 insects each with an insecticide The numbers killed in the various batches were as follows. 30, 54, 33, 60, 63, 48, 54, 63, 54, 51. Does the variance conform to what might be expected on the basis of a binomial distribution of mortality?

2. The following counts of bacteria from the same culture were made on 8 plates: 260, 196, 204, 246, 186, 260, 198, 278. Is the variability in conformity with what might be expected in samples from a Poisson distribution?

3. Test the goodness of fit of the normal curve fitted in exercise 8, page 86

4. Test the goodness of fit of the normal curves fitted in exercise 9, page 86

5. Test the goodness of fit of the normal curves fitted in exercise 10, page 86.

6. Use the χ^2 distribution to test the significance of the difference in death rates in (*a*) exercise 16, page 86; (*b*) exercise 17, page 87; (*c*) exercise 18, page 87.

7. If, in a cross between hybrids, two Mendelian factors are inherited independently, that is, if there is no linkage between the two, then the four possible combinations should theoretically occur in the proportion 9: 3· 3: 1 In an experiment with a species of flower the results shown in Table T were noted. (Gregory, *Journal of Genetics*, vol 1) Are these results consistent with this proportion?

TABLE T

RESULTS OF CROSSING TWO HYBRIDS OF A SPECIES OF FLOWER
(Frequencies of Various Combinations)

Magenta flower Green stigma	Magenta flower Red stigma	Red flower Green stigma	Red flower Red stigma
120	48	36	13

8. Can ıt be concluded from the data of Table U that bottle-feeding is conducıve to malocclusıon of teeth? (Yates, *Supplement to the Journal of the Royal Statistıcal Society*, vol. 1.)

TABLE U

Malocclusion of the Teeth in Infants

	Normal Teeth	Malocclusion
Breast-fed	4	16
Bottle-fed	1	21

CHAPTER VIII

ANALYSIS OF VARIANCE

57. Comparing two variances. Let two independent estimates of the variance of a normally distributed variable X be

$$s_1'^2 = \frac{1}{n_1}\Sigma(X_1 - \bar{X}_1)^2, \quad s_2'^2 = \frac{1}{n_2}\Sigma(X_2 - \bar{X}_2)^2$$

which are based upon n_1 and n_2 degrees of freedom respectively. Then the probability distribution of

$$w = \frac{s_1'^2}{s_2'^2} \tag{1}$$

is known to be *

$$\frac{[(n_1 + n_2 - 2)/2]!\, n_1^{n_1/2} n_2^{n_2/2} w^{(n_1-2)/2} dw}{[(n_1 - 2)/2]!\,[(n_2 - 2)/2]!\,(n_1 w + n_2)^{(n_1+n_2)/2}} \tag{2}$$

and if we make the transformation $w = e^{2z}$ we find that the distribution of

$$z = \frac{1}{2}\log_e w = \frac{1}{2}\log_e \frac{s_1'^2}{s_2'^2} = \log_e \frac{s_1'}{s_2'} \tag{3}$$

is Fisher's distribution †

$$\frac{2[(n_1 + n_2 - 2)/2]!\, n_1^{n_1/2} n_2^{n_2/2} e^{n_1 z} dz}{[(n_1 - 2)/2]!\,[(n_2 - 2)/2]!\,(n_1 e^{2z} + n_2)^{(n_1+n_2)/2}} \tag{4}$$

The subscript 1 *should always be used with the greater estimated variance.*

* See, for example, J. O. Irwin, "Mathematical theorems involved in the analysis of variance," *Journal of the Royal Statistical Society*, vol 94, 1931, pp. 287f.

† See Irwin, *loc. cit* Also see R A. Fisher, "On a distribution yielding the error functions of several well-known statistics," *Proceedings of the International Mathematical Congress*, Toronto, 1924, pp 805–813 In this paper Fisher shows how the normal distribution, the chi-square distribution, and Student's distribution may be regarded as special cases of his general z distribution, and summarizes the chief uses of all these distributions.

The probability that the ratio (1) will be greater than a specified value w is found by integrating (2) from w to ∞. It is more useful, however, to know the value of w for a fixed probability such as 0.01, and values of the ratio w corresponding to probabilities of 0.05 and 0.01 and to various values of n_1 and n_2 have been tabulated by Snedecor.* These values are called respectively the 5 per cent and 1 per cent values of w. Similarly, values of z corresponding to probabilities of 0.05, 0.01, and 0.001 have been worked out by Fisher and Deming and have been published in Fisher's book.† These 5 per cent and 1 per cent values of z are reproduced, by permission, as Tables VII and VIII respectively at the end of this book. *In testing variances Sheppard's correction should not be applied.*

For large values of n_1 and n_2, and for moderate values if they are equal or nearly so, z is approximately normally distributed with standard deviation

$$\sigma_z = \left[\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{\frac{1}{2}} \tag{5}$$

As an illustration, consider two samples composed of 7 and 9 individuals respectively, and having variances 9 6 and 4.8 respectively. Is the variance 9.6 significantly greater than the variance 4.8?

We have

$$n_1 = N_1 - 1 = 6,\ s_1^2 = 9.6,\ s_1'^2 = \frac{N_1 s_1^2}{N_1 - 1} = \frac{7}{6} \times 9.6 = 11.2$$

$$n_2 = N_1 - 1 = 8,\ s_2^2 = 4.8,\ s_2'^2 = \frac{N_2 s_2^2}{N_2 - 1} = \frac{9}{8} \times 4.8 = 5.4$$

$$w = \frac{11.2}{5.4} = 2.074$$

* George W. Snedecor, "Statistical Methods Applied to Experiments in Agriculture and Biology," Collegiate Press, Inc., Ames, Iowa, 1937. These tables are also to be found in C. B Davenport and Merle P Ekas, "Statistical Methods in Biology, Medicine and Psychology" (4th edition), John Wiley & Sons, Inc , New York, 1936. In Snedecor's notation the ratio $s_1'^2/s_2'^2$ is denoted by F

† R A Fisher, "Statistical Methods for Research Workers." These tables are also to be found in Davenport and Ekas, *op. cit.*

From Snedecor's tables we find that the 5 per cent point corresponding to $n_1 = 6$ and $n_2 = 8$ is 3.58 and that the 1 per cent point is 6.37. The obtained value is well within the 5 per cent point, and the first variance can not be regarded as significantly greater than the second, although it is twice as large.

We could also make the test by setting

$$z = \frac{1}{2} \log_e \frac{11\ 2}{5.4} = 1.1513 \log_{10} \frac{11.2}{5.4} = 0.3648$$

From Tables VII and VIII the 5 per cent and 1 per cent values for z corresponding to $n_1 = 6$, $n_2 = 8$ are 0.6378 and 0.9259 respectively, and the obtained value is well inside the 5 per cent point.

Although these two methods are perhaps in most common use, there are other ways of comparing two variances. For example, the transformation

$$w = \frac{n_2 u}{n_1(1 - u)} \quad \text{or} \quad u = \frac{n_1 w}{n_1 w + n_2} \tag{6}$$

will carry the distribution (2) into

$$\frac{[(n_1 + n_2 - 2)/2]!}{[(n_1 - 2)/2]!\,[(n_2 - 2)/2]!} u^{(n_1-2)/2}(1 - u)^{(n_2-2)/2} du \tag{7}$$

whose integral is the incomplete beta function. From tables of this function * the significance of the ratio between two observed variances can be tested.

58. Analysis of variance as applied to linear regression. It is often possible to separate the variance into the constituent parts contributed by various factors, and it is in this possibility that the power and usefulness of Fisher's method known as the *analysis of variance* lie. Let us, for example, consider the regression line fitted in section 18 to the set of points $(X, Y) = (0, 1), (1, 3), (3, 2), (6, 5), (8, 4)$. The equation of this line was found to be $Y' = 1.646 + 0.376X$, and the sum of squares of deviations from it was $\Sigma(Y - Y')^2 = 3.60620$.

* "Tables of the Incomplete Beta-Function," Biometrika Office, London

Now the sum of squares of deviations from the mean can be written

$$\Sigma(Y - \bar{Y})^2 = \Sigma[(Y - Y') + (Y' - \bar{Y})]^2$$
$$= \Sigma(Y - Y')^2 + 2\Sigma(Y - Y')(Y' - \bar{Y}) + \Sigma(Y' - \bar{Y})^2 \quad (8)$$

The middle term is zero, since

$$\Sigma(Y - Y')(Y' - \bar{Y}) = \Sigma(Y - a - bX)(a + bX - \bar{Y})$$
$$= a\Sigma(Y - a - bX) + b\Sigma X(Y - a - bX) - \bar{Y}\Sigma(Y - a - bX)$$

and all three of these terms vanish by reason of the normal equations from which the regression line was determined.

Consequently (8) reduces to

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y')^2 + \Sigma(Y' - \bar{Y})^2 \quad (9)$$

which states that the total variation about the general mean, as measured by the sum of squares of deviations from the mean, is equal to the variation of the residuals about the regression line plus the variation of the regression line about the mean.

The last term in (9) is usually found as the remainder after the variation about the regression line has been subtracted from the total variation. In the present example $\Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/N = 55 - (15)^2/5 = 10$, and $\Sigma(Y' - \bar{Y})^2 = 10 - 3.60620 = 6.39380$. However, the term $\Sigma(Y' - \bar{Y})^2$ can be calculated independently as follows:

$$\Sigma(Y' - \bar{Y})^2 = \Sigma(a + bX - \bar{Y})^2 = \Sigma[\bar{Y} + b(X - \bar{X}) - \bar{Y}]^2$$
$$= b^2\Sigma(X - \bar{X})^2 = \frac{(\Sigma xy)^2}{(\Sigma x^2)^2}\Sigma x^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$
$$= \frac{(\Sigma XY - \Sigma X \cdot \Sigma Y/N)^2}{\Sigma X^2 - (\Sigma X)^2/N} = \frac{(71 - 18 \times 3)^2}{110 - 324/5} = \frac{289}{45.2} = 6.39380$$

There are five points to which the regression line is fitted, so that the number of degrees of freedom for estimating the total variance is four, one having been deducted because we have calculated the mean from the set of values. Another must be deducted for the regression, since another constant, b, has been introduced.

The results may be placed in an analysis of variance table such as Table 24.

TABLE 24

ANALYSIS OF VARIANCE FOR LINEAR REGRESSION

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Residuals	3 60620	3	1 20207
Regression	6 39380	1	6 39380
Total	10	4	

To test the significance of the linear regression we can use the ratio of the two mean square deviations, or half the difference of their natural logarithms,

$$z = \tfrac{1}{2}(\log_e 6.39380 - \log_e 1.20207) = 0.83564$$

Here $n_1 = 1$, $n_2 = 3$, and for these degrees of freedom the 5 per cent and 1 per cent points of z are 1.1577 and 1.7649 respectively. The observed value is within the 5 per cent point, and the regression is not significant.

We have just shown that

$$\Sigma(Y' - \overline{Y})^2 = \Sigma y'^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

We know, moreover, that

$$\Sigma(Y - Y')^2 = \Sigma(y - y')^2 = \Sigma y^2 - b\Sigma xy = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} \quad (10)$$

The sum of these two expressions is the total sum of squares of deviations Σy^2. From the above, and from formula (9), page 49, we find that

$$r^2 = 1 - \frac{\Sigma y^2 - (\Sigma xy)^2/\Sigma x^2}{\Sigma y^2} = \frac{\Sigma y'^2}{\Sigma y^2} \quad (11)$$

Also, since $r^2 = (\Sigma xy)^2/\Sigma x^2 \cdot \Sigma y^2$, we see that we may in general analyze the sum of squares of deviations as follows:

$$\begin{array}{llcrr}
 & & & & \text{Degrees of} \\
 & & & & \text{freedom} \\
\text{Residuals } \Sigma(y-y')^2 & = \Sigma y^2 - \dfrac{(\Sigma xy)^2}{\Sigma x^2} & = & (1-r^2)\Sigma y^2, & N-2 \\[2ex]
\text{Regression } \Sigma y'^2 = b^2\Sigma x^2 = & \dfrac{(\Sigma xy)^2}{\Sigma x^2} & = & r^2\Sigma y^2, & 1 \\
\hline
\text{Total} & \Sigma y^2 & & \Sigma y^2, & N-1
\end{array}$$

Actually, then, testing whether the mean square deviation due to regression is significantly greater than the mean square of residuals incidentally tests whether the correlation coefficient is significant.

The question may be considered from a somewhat more general standpoint. Thus, if we wish to test the significance of the departure of an observed set of data from a hypothetical population regression $Y'_\infty = \alpha + \beta x$ (for simplicity we have taken the origin of x at the mean), we can analyze the sum of squares of deviations from the population regression line into the sum of squares of residual deviations about the sample regression line $Y' = a + bx$ and two other terms as follows:

$$\begin{aligned}
\Sigma(Y - Y'_\infty)^2 &= \Sigma[(Y - Y') + (Y' - Y'_\infty)]^2 \\
&= \Sigma[(Y - a - bx) + (a - \alpha) + (b - \beta)x]^2 \\
&= \Sigma(Y - a - bx)^2 + N(a - \alpha)^2 + (b - \beta)^2\Sigma x^2 \quad (12)
\end{aligned}$$

it being demonstrable that the cross-product terms vanish on summation. The various terms of this equation with their appropriate degrees of freedom are shown in Table 25.

TABLE 25

Analysis of Variance for Linear Regression, Showing Subdivision of Regression Sum of Squares

	Sum of squares of deviations	Degrees of freedom
Residuals	$\Sigma(Y - a - bx)^2$	$N - 2$
Constant term. .	$N(a - \alpha)^2$	1
First degree term .	$(b - \beta)^2\Sigma x^2$	1
Total. . .	$\Sigma(Y - \alpha - \beta x)^2$	N

To test the significance of the first degree term we consider the ratio

$$w = \frac{(b - \beta)^2 \Sigma x^2}{\Sigma(Y - a - bx)^2/(N - 2)} \tag{13}$$

or z, half the natural logarithm of this ratio, and apply the usual test, with $n_1 = 1$, $n_2 = N - 2$.

It will be noted that the square root of (13) is exactly the t of equation (11) of section 43 in Chapter VI. In fact, when $n_1 = 1$, the w or z test and the t test give identical results.

To test the constant term we use the ratio

$$w = \frac{N(a - \alpha)^2}{\Sigma(Y - a - bx)^2/(N - 2)} \tag{14}$$

or the corresponding z.

The test is slightly different from the test of significance of the difference between a mean a and a hypothetical mean α; it tests the significance of the deviation from α of the estimate a of the mean value of Y for a given value of x at or near the sample mean. For a fuller discussion of this point the reader is referred to Fisher, "Statistical Methods for Research Workers," section 26.

To test the significance of the deviation of the regression observed in the sample from the population regression we use

$$w = \frac{[N(a - \alpha)^2 + (b - \beta)^2 \Sigma x^2]/2}{\Sigma(Y - a - bX)^2/(N - 2)} \tag{15}$$

or half its natural logarithm. This can not be changed to a t test because $n_1 = 2$, $n_2 = N - 2$, and the t test can be used only when $n_1 = 1$.

In our illustrative example,

$$N(a - \alpha)^2 = 5(3 - \alpha)^2$$

$$(b - \beta)^2 \Sigma x^2 = (0.376 - \beta)^2 45.2$$

$$\Sigma(Y - a - bX)^2 = 3\ 60620$$

If α and β are zero these quantities assume the values 45, 6.39380 (not exactly, but this is merely because 0.376 has been carried to only three places), and 3.60620, respectively, the last two of which are the values of Table 24.

59. Application to curvilinear and multiple regression and correlation. The results of the preceding section admit of extension to regression of higher order. For instance, we may wish to see whether we get significantly better results, by fitting a second-degree regression curve, over what we get by fitting a straight line. Let the two regression equations be

$$Y' = a_1 + b_1X, \quad Y'' = a_2 + b_2X + c_2X^2 \tag{16}$$

Then it can readily be shown by methods already employed that

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y'')^2 + \Sigma(Y'' - Y')^2 + \Sigma(Y' - \bar{Y})^2 \tag{17}$$

That is, the total variation about the mean may be analyzed into three parts:

(i) the residual variation about the parabola.
(ii) the variation of the parabola about the straight line.
(iii) the variation of the straight line about the mean.

For the set of five points considered in the preceding section, the first- and second-order regression equations were found (see sections 18 and 21) to be

$$Y' = 1\ 646 + 0.376X, \quad Y'' = 1.3460 + 0.73427X - 0.04497X^2$$

respectively. The sums of squares of deviations from these were found to be $\Sigma(Y - Y')^2 = 3.60620$ and $\Sigma(Y - Y'')^2 = 3.22812$. Then $\Sigma(Y'' - Y')^2 = 3.60620 - 3.22812 = 0.37808$.

The analysis of variance table assumes the form below (Table 26). The remaining degree of freedom is taken up by the variation of the mean about the origin.

TABLE 26

Analysis of Variance for Parabolic Regression

	Sum of squares of deviations	Degrees of freedom
Residuals	$\Sigma(Y - Y'')^2 =$ 3 22812	2
Parabola about line	$\Sigma(Y'' - Y')^2 =$ 0 37808	1
Line about mean . . .	$\Sigma(Y' - \bar{Y})^2 =$ 6 39380	1
Total . ..	$\Sigma(Y - \bar{Y})^2 =$ 10 00000	4

To test the significance of the deviation of the parabola from the straight line we take the ratio

$$w = \frac{0\ 37808}{3.22812/2} = \frac{0.37808}{1.61406} = 0.23424$$

The degrees of freedom are $n_1 = 1$, $n_2 = 2$. Three possible ways of completing the test are now open to us.* We may use Snedecor's tables on w directly, or Tables VII and VIII on $z = \frac{1}{2} \log_e 0.23424$. A third method, which is the one we shall use, is to calculate $t = w^{\frac{1}{2}} = 0.484$. For two degrees of freedom this is decidedly not significant, since the probability of a value of t this large or larger is greater than 0.6.

To test the significance of the linear part of the regression we can compute

$$t = w^{\frac{1}{2}} = \left(\frac{6.39380}{1.61406}\right)^{\frac{1}{2}} = (3.9613)^{\frac{1}{2}} = 1.990, \quad n = 2$$

The probability of a value of t numerically this large or larger is between 0.2 and 0.1. Although this is not significant, it is less probable than the value obtained from $t = 0.484$. In some cases it is possible in this way to show, for example, that the linear regression fitted to a set of data is significant while a second-order regression is not, that is, that nothing is to be gained in such cases by fitting anything but a linear regression.

To test the complete regression in our illustration we could use

$$w = \frac{(6\ 39380 + 0.37808)/2}{3.22812/2} = 2.0978$$

or

$$z = \tfrac{1}{2} \log_e w = 0.3704$$

Here we have $n_1 = 2$, $n_2 = 2$, and we can not use the t test. The values are not significant.

For a multiple regression equation such as

$$Y' = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

we can form Table 27. The sums of squares of deviations have been expressed in terms of the multiple correlation coefficient† R. (See section 28.)

* Actually there is no necessity of making the test, since $w < 1$.

† For convenience we use R instead of $r_{1\ 23\ \cdot k}$

TABLE 27

ANALYSIS OF VARIANCE FOR MULTIPLE REGRESSION

	Sum of squares of deviations	Degrees of freedom
Residuals	$\Sigma(Y - Y')^2 = (1 - R^2)\Sigma(Y - \bar{Y})^2$	$N - k - 1$
Regression	$\Sigma(Y' - \bar{Y})^2 = R^2\Sigma(Y - \bar{Y})^2$	k
Total .	$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - \bar{Y})^2$	$N - 1$

To test the significance of the multiple regression we form the ratio

$$w = \frac{\Sigma(Y' - \bar{Y})^2/k}{\Sigma(Y - Y')^2/(N - k - 1)} = \frac{R^2/k}{(1 - R^2)/(N - k - 1)} \tag{18}$$

and use Snedecor's tables with $n_1 = k$, $n_2 = N - k - 1$, or set $z = \frac{1}{2} \log_e w$ and use Tables VII and VIII with these same degrees of freedom. Incidentally we are testing the significance of the multiple correlation coefficient R.

60. Absolute criteria in the theory of regression.* In the preceding section, our estimate of the variance of Y, without utilizing the regression, is $\Sigma(Y - \bar{Y})^2/(N - 1)$. Our estimate of the variance of Y taking the regression into consideration is $(1 - R^2)\Sigma(Y - \bar{Y})^2/(N - k - 1)$. Consequently if our estimation is to be improved by the use of regression, that is, if our estimate is to have a smaller variance, we must have

$$\frac{(1 - R^2)\Sigma(Y - \bar{Y})^2}{N - k - 1} < \frac{\Sigma(Y - \bar{Y})^2}{N - 1} ,$$

which leads to the inequality

$$R^2 > k/(N - 1) \tag{19}$$

To be somewhat more exact, our estimate of the variance of an individual forecast without using the regression is our estimate of

* From an unpublished note by Churchill Eisenhart.

the variance about the mean plus the variance of the mean itself. This estimate is

$$\frac{\Sigma(Y - \overline{Y})^2}{N - 1}\left(1 + \frac{1}{N}\right) \tag{20}$$

The minimum value of the estimate of the variance of a forecast when the regression is used occurs when $X_1 = \overline{X}_1, \ldots, X_k = \overline{X}_k$, and is equal to *

$$\frac{(1 - R^2)\Sigma(Y - \overline{Y})^2}{N - k - 1}\left(1 + \frac{1}{N}\right) \tag{21}$$

If (21) is to be less than (20), then

$$\frac{1 - R^2}{N - k - 1} < \frac{1}{N - 1}$$

which leads to (19) as before. For $k = 1$, (19) degenerates into the inequality

$$r^2 > \frac{1}{N - 1} \tag{22}$$

for the usual correlation coefficient associated with linear regression.

By using such inequalities as the above, we can make a definite decision regarding the advisability of keeping an additional independent variable in our regression. Let R_k^2 denote the square of the coefficient of multiple correlation for k independent variables, and R_{k+1}^2 the square of the coefficient for $k + 1$ independent variables. If the additional variable is to improve our estimation, we must have

$$\frac{1 - R_{k+1}^2}{N - k - 2} < \frac{1 - R_k^2}{N - k - 1}$$

or

$$\frac{1 - R_{k+1}^2}{1 - R_k^2} < 1 - \frac{1}{N - k - 1} \tag{23}$$

This is probably the most convenient form of the inequality, as it gives the necessary smallness of the ratio of the two residual sums

* Cf. R A Fisher, "Statistical Methods for Research Workers," section 26

of squares if the additional variable is to improve our estimation. Thus, if a simple linear regression ($k = 1$) has been calculated on twelve observations, we see that it is useless to consider the addition of a second independent variable unless by doing so we can reduce the residual sum of squares by more than

$$\frac{1}{N-k-1} = \frac{1}{12-1-1} = 10 \text{ per cent}$$

It is not intended to suggest that these criteria be used as tests of significance, but merely to indicate that there exist such absolute criteria—criteria independent of significance level chosen—and to suggest that these considerations be taken into account in experiments in which the use of simple or multiple regression methods may be contemplated.

61. Testing the significance of the correlation ratio. In section 25 two different forms for the square of the correlation ratio were given, viz,

$$\eta^2 = 1 - \sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{Xi} - \bar{Y}_X)^2 \Big/ \sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{Xi} - \bar{Y})^2 \quad (24)$$

and

$$\eta^2 = \sum_{X=X_1}^{X_k} N_X(\bar{Y}_X - \bar{Y})^2 \Big/ \sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{Xi} - \bar{Y})^2 \quad (25)$$

For the meaning of the notation, refer to section 25. If we equate these two values and clear of fractions we obtain the fundamental identity

$$\sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{Xi} - \bar{Y})^2$$

$$= \sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{Xi} - \bar{Y}_X)^2 + \sum_{X=X_1}^{X_k} N_X(\bar{Y}_X - \bar{Y})^2 \quad (26)$$

which expresses the fact that in a correlation table the sum of squares of deviations of the Y's about their mean is equal to the sum of squares of the deviations about the means of columns plus the (weighted) sum of squares of the deviations of the means of columns about the general mean. This enables us to form the analysis of variance table (Table 28).

TABLE 28

Analysis of Variance for Correlation Ratio

	Sum of squares of deviations	Degrees of freedom
Within columns	$\sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{X_i}-\bar{Y}_X)^2=(1-\eta^2)\sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{X_i}-\bar{Y})^2$	$N-k$
Column means	$\sum_{X=X_1}^{X_k}N_X(\bar{Y}_X-\bar{Y})^2 = \eta^2\sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{X_i}-\bar{Y})^2$	$k-1$
Total	$\sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{X_i}-\bar{Y})^2 = \sum_{X=X_1}^{X_k}\sum_{i=1}^{N_X}(Y_{X_i}-\bar{Y})^2$	$N-1$

$$N=\sum_{X=X_1}^{X_k}N_X$$

We can calculate the ratio of the mean square deviation of the column means to the mean square deviation within the columns,

$$w=\frac{\eta^2/(k-1)}{(1-\eta^2)/(N-k)}=\frac{N-k}{k-1}\cdot\frac{\eta^2}{1-\eta^2} \tag{27}$$

and use either w or ½ $\log_e w$ in the usual manner to test the significance of an observed value of η.

As an illustration we shall consider the value of η computed in section 25 of Chapter IV. There we found

$$\Sigma_X\Sigma_i(Y_{X_i}-\bar{Y}_X)^2=1+2.83+4.4+2+0=10.23$$

$$\Sigma_X N_X(\bar{Y}_X-\bar{Y})^2=5.77,\quad \Sigma_X\Sigma_i(Y_{X_i}-\bar{Y})^2=16$$

The number of columns was $k = 5$, and the total number of items in the correlation table was $N = 25$. Consequently the analysis of variance table assumes the form shown in Table 29.

TABLE 29

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Within columns	10 23	$N - k = 20$	0 5165
Column means	5 77	$k - 1 = 4$	1 4425
Total	16	$N - 1 = 24$	

$$z = \frac{1}{2}\log_e \frac{1\ 4425}{0\ 5165} = 0\ 5147, \quad n_1 = 4, \quad n_2 = 20$$

This is just under the 5 per cent point 0.5265 and can hardly be regarded as significant.

62. Testing linearity of regression. To test linearity of regression, that is, to test whether η^2 is significantly larger than r^2, we subdivide the sums of squares still further.

Suppose that in a correlation table we have fitted a regression line $Y' = a + bX$. Let Y'_X be the ordinate of this line corresponding to a fixed value of X, and let $\bar{Y}_X$ be the mean of Y for this value of X, that is, the mean of the column whose central abscissa is X. Then it can be demonstrated that

$$\sum_X \sum_i (Y_{Xi} - \bar{Y})^2 = \sum_X \sum_i (Y_{Xi} - \bar{Y}_X)^2 + \sum_X N_X(\bar{Y}_X - Y'_X)^2 + \sum_X N_X(Y'_X - \bar{Y})^2 \quad (28)$$

In words, the sum of squares of deviations about the mean is composed of the sum of squares of deviations about the column means, the (weighted) sum of squares of the deviations of the column means from the regression line, and the (weighted) sum of squares of deviations of the ordinates of the regression line about the general mean.

The sum of the first two terms on the right of (28) is equal to

$$\Sigma_X\Sigma_i(Y_{Xi} - Y'_X)^2$$

which is the sum of squares of deviations about the straight line of regression, and which is consequently equal to $(1 - r^2)\Sigma_X N_X(\overline{Y}_X - \overline{Y})^2$. It is not difficult to prove the other relations necessary for setting up the analysis of variance table (Table 30).

TABLE 30

ANALYSIS OF VARIANCE FOR TESTING LINEARITY OF REGRESSION

	Sum of squares of deviations	Degrees of freedom
Residuals about column means	$\sum_X\sum_i(Y_{X_i}-\overline{Y}_X)^2 = (1-\eta^2)\sum_X N_X(\overline{Y}_X-\overline{Y})^2$	$N - k$
Column means about regression line	$\sum_X N_X(\overline{Y}_X-Y'_X)^2 = (\eta^2-r^2)\sum_X N_X(\overline{Y}_X-\overline{Y})^2$	$k - 2$
About regression line Subtotal .	$\sum_X\sum_i(Y_{X_i} - Y'_X)^2 = (1-r^2)\sum_X N_X(\overline{Y}_X-\overline{Y})^2$	$N - 2$
Regression .	$\sum_X N_X(Y'_X-\overline{Y})^2 = r^2\sum_X N_X(\overline{Y}_X-\overline{Y})^2$	1
Total	$\sum_X\sum_i(Y_{X_i}-\overline{Y})^2 = \sum_X\sum_i(Y_{X_i}-\overline{Y})^2$	$N - 1$

To test whether η is significantly greater than r we set

$$z = \frac{1}{2}\log_e \frac{\Sigma_X N_X(\overline{Y}_X - Y'_X)^2/(k-2)}{\Sigma_X\Sigma_i(Y_{X_i} - \overline{Y}_X)^2/(N-k)}$$

$$= \frac{1}{2}\log_e \frac{(\eta^2 - r^2)/(k-2)}{(1-\eta^2)/(N-k)} \tag{29}$$

$$n_1 = k - 2, \quad n_2 = N - k$$

and proceed in the usual manner.

In the example already cited,

$$\eta^2 = 0.3604, \quad r^2 = 0.3299, \quad N = 25, \quad k = 5 \text{ (No. of columns)}$$

$$z = \frac{1}{2}\log_e \frac{(0.3604 - 0.3299)/3}{(1 - 0.3604)/20} = \frac{1}{2}\log_e \frac{0.01017}{0.03198}$$

The numerator of the ratio is smaller than the denominator and consequently the difference between η^2 and r^2 can not possibly be significant.

63. Variance within and among classes. Suppose that we have k classes of individuals with m individuals in each class. We may wish to discover whether there is significantly more variation among the classes, that is, among the class means, than there is within the classes. Let the measurement of the characteristic of the ith individual in the jth class be denoted by X_{ij}, and let the mean of the jth class be denoted by $\overline{X}_j$. (See Table 31.) Then

TABLE 31

	Class					
	1	2	·	j	·	k
	X_{11} X_{21} · · · X_{i1} X_{m1}	X_{12} X_{22} · X_{i2} · X_{m2}	·· ··· · · ·	X_{1j} X_{2j} · X_{ij} · X_{mj}	 ·· · · · ·	X_{1k} X_{2k} · X_{ik} X_{mk}
Total	T_1	T_2	·	T_j	··	T_k
Mean	$\overline{X}_1$	$\overline{X}_2$	· ·	$\overline{X}_j$	·	$\overline{X}_k$

$$\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\overline{X})^2=\sum_{j=1}^{k}\sum_{i=1}^{m}[(X_{ij}-\overline{X}_j)+(\overline{X}_j-\overline{X})]^2$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\overline{X}_j)^2+2\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\overline{X}_j)(\overline{X}_j-\overline{X})$$

$$+\sum_{j=1}^{k}\sum_{i=1}^{m}(\overline{X}_j-\overline{X})^2$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\overline{X}_j)^2+m\sum_{j=1}^{k}(\overline{X}_j-\overline{X})^2 \qquad (30)$$

the middle term vanishing on summation. The importance of this formula lies in the fact that it separates the sum of squares of deviations about the general mean into the sum of squares of deviations about the class means and the sum of squares of deviations of the class means about the general mean (multiplied by the number in each class, of course). These two components may be spoken of as the sum of squares of deviations *within classes* and *among classes* respectively. The formula is exhibited in tabular form in Table 32. The total number of items in the entire group is mk, and the total number of degrees of freedom is $mk - 1$, unity having been deducted because we are calculating deviations about the mean. That is, one degree of freedom is taken up by the deviation of the mean about the zero point. The number of degrees of freedom within each class is similarly $m - 1$, and since there are k classes, the number of degrees of freedom within classes is $k(m - 1)$. The number of degrees of freedom among classes is $k - 1$, one less than the number of classes.

TABLE 32

ANALYSIS OF VARIANCE WITHIN AND AMONG CLASSES

	Sum of squares of deviations	Degrees of freedom
Within classes	$\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij} - \bar{X}_j)^2$	$k(m - 1)$
Among classes	$m\sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$k - 1$
Total	$\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij} - \bar{X})^2$	$mk - 1$

To test whether the variance among classes is significant we use

$$w = \frac{m\Sigma_j(\bar{X}_j - \bar{X})^2/(k - 1)}{\Sigma_j\Sigma_i(X_{ij} - \bar{X}_j)^2/k(m - 1)} \tag{31}$$

or $z = \frac{1}{2}\log w$, with $n_1 = k - 1$, $n_2 = k(m - 1)$.

For actual computation it is probably preferable to transform the foregoing formulas somewhat. Thus, for the total sum of squares of deviations we have

$$\begin{aligned}\Sigma_j\Sigma_i(X_{ij} - \bar{X})^2 &= \Sigma_j\Sigma_i X_{ij}^2 - \frac{(\Sigma_j\Sigma_i X_{ij})^2}{N} \\ &= \Sigma_j\Sigma_i X_{ij}^2 - \frac{T^2}{N} \qquad (32)\end{aligned}$$

where $N = mk$, and T is the grand total.

For the sum of squares of deviations among the class means we find

$$\begin{aligned}m\Sigma_j(\bar{X}_j - \bar{X})^2 &= m\Sigma_j\bar{X}_j^2 - \frac{(\Sigma_j\Sigma_i X_{ij})^2}{N} \\ &= m\Sigma_j\left(\frac{\Sigma_i X_{ij}}{m}\right)^2 - \frac{1}{N}(\Sigma_j\Sigma_i X_{ij})^2 \\ &= \frac{\Sigma_j T_j^2}{m} - \frac{T^2}{N} \qquad (33)\end{aligned}$$

T_j being the total of the jth class.

The sum of squares of deviations within classes can be calculated as a remainder by subtracting the sum of squares of deviations among class means from the total sum of squares of deviations.

It may be remarked that

$$w = \frac{1 + (m - 1)r'}{1 - r'} \qquad (34)$$

where r' is the *intraclass correlation coefficient*,* so that our test will incidentally test the significance of this coefficient.

As an illustration of variance among classes let us consider Table 33, which shows the ulna lengths of four strains of hens. Measurements of five individuals from each strain are given. Actually, measurements would doubtless be made to a greater degree of accuracy. These numbers are given to only two significant figures for simplicity of illustration.

* See R. A. Fisher, "Statistical Methods for Research Workers."

TABLE 33

ULNA LENGTHS (IN MILLIMETERS) IN 4 STRAINS OF HENS

	Strain			
	1	2	3	4
	67	68	72	66
	66	68	70	70
	66	71	68	65
	73	69	65	64
	66	58	70	67
Total . .	338	334	345	332
Mean	67.6	66.8	69.0	66.4

$$\text{Grand total} = T = 1349$$

$$\text{General mean} = \bar{X} = \frac{1349}{20} = 67.45$$

Ordinarily the calculation would be performed on a machine.* In the present example we shall list the squares of the items in Table 34 so that the application of the formulas involved in the analysis can be more easily understood.

TABLE 34

SQUARES OF ITEMS IN TABLE 33

4,489	4,624	5,184	4,356
4,356	4,624	4,900	4,900
4,356	5,041	4,624	4,225
5,329	4,761	4,225	4,096
4,356	3,364	4,900	4,489
22,886	22,414	23,833	22,066

Grand total = 91,199

* It might sometimes be of advantage to express all numbers as deviations from an arbitrary origin. For example, if we had a long series of numbers in the 700's, such as 725, 702, 718, . . , we could write 25, 2, 18, . . , and, dealing with these deviations, obtain precisely the same results

From formula (32) we find for the total sum of squares of deviations

$$91{,}199 - \frac{(1349)^2}{20} = 91{,}199 - 90{,}990.05 = 208.95$$

From (33) the sum of squares of deviations among means of strains is

$$\tfrac{1}{5}[(338)^2 + (334)^2 + (345)^2 + (332)^2] - \frac{(1349)^2}{20}$$

$$= \tfrac{1}{5}(114{,}244 + 111{,}556 + 119{,}025 + 110{,}224) - \frac{1{,}819{,}801}{20}$$

$$= \tfrac{1}{5} \times 455{,}049 - 90{,}990.05 = 19.75$$

The sum of squares of deviations within strains is the difference 208.95 − 19.75 = 189.20, and the analysis of variance table is as follows:

TABLE 35

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Among strains.	19 75	$4 - 1 = 3$	6 583
Within strains	189 20	$4(5 - 1) = 16$	11.8
Total .	208.95	$4 \times 5 - 1 = 19$	

It is seen that there is less variation among strains than within strains.

If there are unequal numbers in the various classes, say N_1 in the first class, N_2 in the second, and so on, the fundamental formula is

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(X_{ij} - \bar{X})^2$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(X_{ij} - \bar{X}_j)^2 + \sum_{j=1}^{k} N_j(\bar{X}_j - \bar{X})^2 \qquad (35)$$

This is exactly. comparable to the formula (26) used in connection with the correlation ratio. In fact, our classes correspond to the columns of the correlation table from which the correlation ratio is calculated, and consequently to test whether the variation among class means is significantly greater than within classes we proceed as before, using the ratio

$$w = \frac{\Sigma_j N_j(\bar{X}_j - \bar{X})^2/(k - 1)}{\Sigma_j \Sigma_i (X_{ij} - \bar{X}_j)^2/(N - k)} \tag{36}$$

or half its natural logarithm.

64. Subdivision of variance into more than two portions.* It often happens that there is a connection between the individual items in the different classes. For example, we might have observations on the first flowering date of four different varieties of plants at ten different stations. The observations could be classified both according to variety and to locality. We should in general have a table like Table 31, but should want the means of rows as well as columns. (See Table 36.)

TABLE 36

	1	2 · ·	j · · ·	k	Mean
1	X_{11}	X_{12} · · ·	X_{1j} · · ·	X_{1k}	$\bar{X}_1$
2	X_{21}	X_{22} · · ·	X_{2j} · · ·	X_{2k}	$\bar{X}_2$
· · ·	· ·	· · ·	· · ·	·	·
i	X_{i1}	X_{i2} · · ·	X_{ij} · · ·	X_{ik}	$\bar{X}_i$
· · ·	· ·	· · ·	· · ·	·	· · ·
m	X_{m1}	X_{m2} · ·	X_{mj} · · ·	X_{mk}	$\bar{X}_{m\cdot}$
Mean	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$ · ·	$\bar{X}_{\cdot j}$ · ·	$\bar{X}_{\cdot k}$	$\bar{X}$

[*] To obtain a clear insight into the foundations of the analysis of variance one should read two papers by J. O. Irwin, "Mathematical theorems involved in the analysis of variance," *Journal of the Royal Statistical Society*, vol. 94, 1931, pp 284–300; and "Independence of the constituent items in the analysis of variance," *Supplement to the Journal of the Royal Statistical Society*, vol. 1, 1934, pp. 236–251.

Let $\bar{X}_{i\cdot}$ denote the mean of the ith row, $\bar{X}_{\cdot j}$ the mean of the jth column, and $\bar{X}$ the general mean. The fundamental identity is

$$\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\bar{X})^2 = k\sum_{i=1}^{m}(\bar{X}_{i\cdot}-\bar{X})^2 + m\sum_{j=1}^{k}(\bar{X}_{\cdot j}-\bar{X})^2$$
$$+\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X})^2 \quad (37)$$

The last term is called the *error* or *interaction* It has been freed of the effect of both rows and columns, and is therefore assumed to be due to experimental error only Ordinarily it is not calculated directly, but as the remainder after deducting the other two terms from the total sum of squares of deviations. Exhibiting formula (37) as an analysis of variance table, we have Table 37. The number of degrees of freedom for rows (or columns) is one less than the number of rows (or columns). The number of degrees of freedom for error is, as in a contingency table, the number of compartments of the table that can be arbitrarily filled, keeping the marginal totals or means constant, viz., $(m - 1)(k - 1)$.

TABLE 37

	Sum of squares of deviations	Degrees of freedom
Means of rows	$k\Sigma_i(\bar{X}_i - \bar{X})^2$	$m - 1$
Means of columns	$m\Sigma_j(\bar{X}_{\ j} - \bar{X})^2$	$k - 1$
Error .	$\Sigma_j\Sigma_i(X_{ij} - \bar{X}_i - \bar{X}_{\ j} + \bar{X})^2$	$(m - 1)(k - 1)$
Total .	$\Sigma_j\Sigma_i(X_{ij} - \bar{X})^2$	$mk - 1$

If our data are a random sample from a homogeneous normal population, then each sum of squares of deviations in Table 37, when divided by the corresponding number of degrees of freedom, gives an unbiased estimate of the variance of the population. If the population is not homogeneous, that is, if it is more variable in one way than another, this will show up in the different estimates of the population variance; some of them will be greater than others. It is usual to compare the columns and rows with the error or interaction.

To test whether there is significant variation in the means of rows we use

$$w = \frac{k\Sigma_i(\bar{X}_i - \bar{X})^2/(m-1)}{\Sigma_j\Sigma_i(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\,j} + \bar{X})^2/(m-1)(k-1)} \tag{38}$$

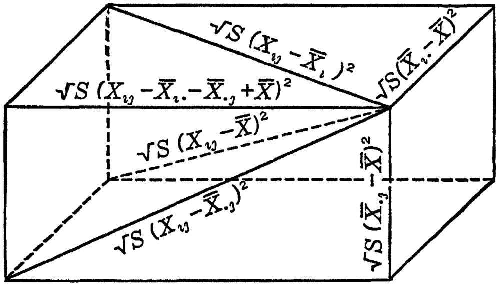
$$n_1 = m - 1, \quad n_2 = (m-1)(k-1)$$

and to test the variation of the means of columns we use

$$w = \frac{m\Sigma_j(\bar{X}_{\,j} - \bar{X})^2/(k-1)}{\Sigma_j\Sigma_i(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2/(m-1)(k-1)} \tag{39}$$

$$n_1 = k - 1, \quad n_2 = (m-1)(k-1)$$

Of course, we have the option of using $z = \frac{1}{2}\log_e w$ in either case.



$\sqrt{S}\,(X_{ij} - \bar{X}_{i})^2$
$\sqrt{S}(\bar{X}_{i\cdot} - \bar{X})^2$
$\sqrt{S}\,(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$
$\sqrt{S}\,(X_{ij} - \bar{X})^2$
$\sqrt{S}\,(\bar{X}_{\cdot j} - \bar{X})^2$
$\sqrt{S}\,(X_{ij} - \bar{X}_{\cdot j})^2$

FIG 11—Analysis of Variance Diagram.

Formula (37) can be represented geometrically by letting the square root of the left side be the diagonal of a rectangular parallelopiped of which the edges are the square roots of the three terms on the right side, as in Fig. 11. In this figure the symbol S is used to indicate summation over every individual in the sample; e.g.,

$$S(\bar{X}_{i\cdot} - \bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X})^2 = k\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X})^2$$

As a very simple illustration of formula (37) let us consider Table 38. We find: for total

TABLE 38

					Total	Mean
	2 4 6	3 7 5	5 8 5	6 5 4	16 24 20	4 6 5
Total	12	15	18	15	60	15
Mean.	4	5	6	5	20	5

$$\sum_{j=1}^{4}\sum_{i=1}^{3}(X_{ij}-\bar{X})^2 = (2-5)^2+(3-5)^2+(5-5)^2+(6-5)^2$$
$$+(4-5)^2+(7-5)^2+(8-5)^2+(5-5)^2$$
$$+(6-5)^2+(5-5)^2+(5-5)^2+(4-5)^2$$
$$= 9+4+0+1+1+4+9+0+1+0+0+1 = 30$$

For means of rows

$$4\sum_{i=1}^{3}(\bar{X}_{i\cdot}-\bar{X})^2$$
$$= 4[(4-5)^2+(6-5)^2+(5-5)^2] = 4(1+1+0) = 8$$

For means of columns

$$3\sum_{j=1}^{4}(\bar{X}_{\cdot j}-\bar{X})^2 = 3[(4-5)^2+(5-5)^2+(6-5)^2+(5-5)^2]$$
$$= 3(1+0+1+0) = 6$$

$$\sum_{j=1}^{4}\sum_{i=1}^{3}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X})^2$$
$$= (2-4-4+5)^2+(3-4-5+5)^2+(5-4-6+5)^2$$
$$+(6-4-5+5)^2+(4-6-4+5)^2+(7-6-5+5)^2$$
$$+(8-6-6+5)^2+(5-6-5+5)^2+(6-5-4+5)^2$$
$$+(5-5-5+5)^2+(5-5-6+5)^2+(4-5-5+5)^2$$
$$= 1+1+0+4+1+1+1+1+4+0+1+1 = 16$$

Check: $30 = 8 + 6 + 16$

In actual practice the sums of squares of deviations would not be calculated as above. Perhaps the most satisfactory formula for use in machine calculation is the following:

$$\Sigma_j\Sigma_i(X_{ij} - \bar{X})^2 = \Sigma_j\Sigma_i X_{ij}^2 - \frac{(\Sigma_j\Sigma_i X_{ij})^2}{N} = \Sigma_j\Sigma_i X_{ij}^2 - \frac{T^2}{N} \qquad (40)$$

in which $N = mk$, and T is the grand total.

In the present example this total sum of squares of deviations would be

$$2^2+3^2+5^2+6^2+4^2+7^2+8^2+5^2+6^2+5^2+5^2+4^2- \frac{(60)^2}{12}$$

$$= 330 - 300 = 30$$

The sum of squares, 300, and the total, 60, could be run off simultaneously on the machine.

Then for rows we could use the formula

$$k\Sigma_i(\bar{X}_{i\cdot} - \bar{X})^2 = k\Sigma_i\bar{X}_{i\cdot}^2 - \frac{(\Sigma_j\Sigma_i X_{ij})^2}{N}$$

$$= k\Sigma_i\left(\frac{\Sigma_j X_{ij}}{k}\right)^2 - \left(\frac{\Sigma_j\Sigma_i X_{ij}}{N}\right)^2 = \frac{1}{k}\,\Sigma_i\,(\Sigma_j X_{ij})^2 - \frac{1}{N}\,(\Sigma_j\Sigma_i X_{ij})^2$$

$$= \frac{\Sigma_i T_i^2}{k} - \frac{T^2}{N} \qquad (41)$$

where T_i is the total of the ith row, and T is the grand total. This gives

$$\tfrac{1}{4}[(16)^2 + (24)^2 + (20)^2] - \tfrac{1}{12}\,(60)^2 = 308 - 300 = 8$$

The formula for columns, analogous to (41), is

$$m\Sigma_j(\bar{X}_{\cdot j} - \bar{X})^2 = \frac{1}{m}\,\Sigma_j(\Sigma_i X_{ij})^2 - \frac{1}{N}\,(\Sigma_j\Sigma_i X_{ij})^2$$

$$= \frac{\Sigma_j T_j^2}{m} - \frac{T^2}{N}$$

$$= \tfrac{1}{3}[(12)^2+(15)^2+(18)^2+(15)^2] - \tfrac{1}{12}(60)^2 = 306-300 = 6$$

T_j is the total of column j.

As has been stated above, the error term would be obtained by subtraction (30 − 8 − 6 = 16).

The analysis of variance table would appear as follows:

TABLE 39

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Rows	8	$3 - 1 = 2$	4
Columns	6	$4 - 1 = 3$	2
Error	16	$3 \times 2 = 6$	$2\ \dot{6}$
Total	30	$3 \times 4 - 1 = 11$	

To test the significance of the variation among rows we set

$$z = \frac{1}{2}\log_e w = \frac{1}{2}\log_e \frac{4}{2\ \dot{6}} = \frac{1}{2}\log_e 1.5 = 0.2027$$

$$n_1 = 2, \quad n_2 = 6$$

Such a value would not be significant.

The mean square deviation for columns is less than that for error and would not need to be tested.

As was suggested above, the observations might be first flowering dates of plants. The columns might then refer to different varieties and the rows to different observation localities. Or the columns might refer to different varieties and the rows to different years. For rainfall data we might have twelve columns for the various months and use the rows for years. Other examples could be adduced without limit, and the method is very broadly useful.

The foregoing method of subdivision admits of still further extension. Rainfall data, for instance, might be classified according to month and year and station at which measured. Or for a given station it might be classified according to the hour of the day as well as according to month and year.

Suppose then that we have $N = klm$ observed values which are subjected to a triple classification into groups, columns, and rows. (See Table 40.) Let there be l groups of k rows and m columns each. Let X_{hij} be the item in row h and column j of the ith group. Let

TABLE 40

	Column			Mean
	1	j	m	
	X_{111}	X_{11j}	X_{11m}	$\bar{X}_{11\cdot}$
	. .	. .	. .	. . .
Group 1	X_{h11}	X_{h1j}	X_{h1m}	$\bar{X}_{h1\cdot}$
	. . .	. . .	. . .	. . .
	X_{k11}	X_{k1j}	X_{k1m}	$\bar{X}_{k1}$
Mean .	$\bar{X}_{\ 11}$	$\bar{X}_{\cdot 1j}$	$\bar{X}_{\ 1m}$	$\bar{X}_{\ 1\cdot}$
. . . .	. . . .	. . . .	. . . .	. . . .
	X_{1i1}	X_{1ij}	X_{1im}	$\bar{X}_{1i\cdot}$
	. . .	. . .	. . .	. .
Group i	X_{hi1}	X_{hij}	X_{him}	$\bar{X}_{hi\cdot}$
	. .	.	. . .	. . .
	X_{ki1}	X_{kij}	X_{kim}	$\bar{X}_{ki\cdot}$
Mean	$\bar{X}_{\cdot i1}$	$\bar{X}_{\ ij}$	$\bar{X}_{\ im}$	$\bar{X}_{\ i\cdot}$
. . . .	. . . .	. . . .	. . . .	. . . .
	X_{1l1}	X_{1lj}	X_{1lm}	$\bar{X}_{1l\cdot}$
	. .	. . .	. . .	. . .
Group l	X_{hl1}	X_{hlj}	X_{hlm}	$\bar{X}_{hl\cdot}$
	. . .	. . .	. . .	. . .
	X_{kl1}	X_{klj}	X_{klm}	$\bar{X}_{kl\cdot}$
Mean	$\bar{X}_{\ l1}$	$\bar{X}_{\ lj}$	$\bar{X}_{\ lm}$	$\bar{X}_{\ l}$
	$\bar{X}_{1\ 1}$	$\bar{X}_{1\cdot j}$	$\bar{X}_{1\ m}$	$\bar{X}_{1\cdot\cdot}$
	. .	. . .	. .	. .
Means of rows	$\bar{X}_{h\ 1}$	$\bar{X}_{h\ j}$	$\bar{X}_{h\ m}$	$\bar{X}_{h\cdot\cdot}$
	. . .	.	. . .	. . .
	$\bar{X}_{k\ 1}$	$\bar{X}_{k\ j}$	$\bar{X}_{k\ m}$	$\bar{X}_{k\cdot\cdot}$
Mean	$\bar{X}_{\ \ 1}$	$\bar{X}_{\ \ j}$	$\bar{X}_{\ \ m}$	$\bar{X}$

$\bar{X}_{hi\cdot}$ = mean of items in row h of group i.

$\bar{X}_{h\cdot j}$ = mean of items in row h and column j.
$\bar{X}_{\cdot ij}$ = mean of items in column j of group i.
$\bar{X}_{h}$ = mean of ıtems ın all hth rows.
$\bar{X}_{\cdot i \cdot}$ = mean of items in group i.
$\bar{X}_{\cdot\cdot j}$ = mean of items ın all jth columns.
$\bar{X}$ = general mean.

Then the fundamental identity, which we shall state without proof, ıs

$$\sum_{j=1}^{m}\sum_{i=1}^{l}\sum_{h=1}^{k}(X_{hij}-\bar{X})^2 = lm\sum_{h=1}^{k}(\bar{X}_{h\cdot\cdot}-\bar{X})^2$$
$$+\,km\sum_{i=1}^{l}(\bar{X}_{\cdot i\cdot}-\bar{X})^2+kl\sum_{j=1}^{m}(\bar{X}_{\cdot\cdot j}-\bar{X})^2$$
$$+\,m\sum_{h=1}^{k}\sum_{i=1}^{l}(\bar{X}_{hi\cdot}-\bar{X}_{h\cdot\cdot}-\bar{X}_{\cdot i\cdot}+\bar{X})^2$$
$$+\,l\sum_{h=1}^{k}\sum_{j=1}^{m}(\bar{X}_{h\cdot j}-\bar{X}_{h\cdot\cdot}-\bar{X}_{\cdot\cdot j}+\bar{X})^2$$
$$+\,k\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{X}_{\cdot ij}-\bar{X}_{\cdot i\cdot}-\bar{X}_{\cdot\cdot j}+\bar{X})^2$$
$$+\sum_{h=1}^{k}\sum_{i=1}^{l}\sum_{j=1}^{m}(X_{hij}-\bar{X}_{hi\cdot}-\bar{X}_{h\cdot j}-\bar{X}_{\cdot ij}+\bar{X}_{h\cdot\cdot}$$
$$+\,\bar{X}_{\cdot i\cdot}+\bar{X}_{\cdot\cdot j}-\bar{X})^2 \qquad (42)$$

The corresponding analysis of variance is shown in Table 41.

The matter will be much clearer if we work through a concrete example. Let us consider Table 42, which shows the first flowering date (day of the year) of 5 varıeties of plants at 6 different stations over a period of 3 years. In the preceding terminology, the plants correspond to columns, the stations to groups, and the years to rows. In this table totals rather than means are shown.

The total sum of squares of deviations is found by summing the squares of all individual items and then subtracting the square of

TABLE 41

ANALYSIS OF VARIANCE FOR TRIPLE CLASSIFICATION (m COLUMNS, l GROUPS, k ROWS)

	Sum of squares of deviations	Degrees of freedom
Means of rows	$lm \sum_{h=1}^{k} (\bar{X}_{h\cdot\cdot} - \bar{X})^2$	$k - 1$
Means of groups	$km \sum_{i=1}^{l} (\bar{X}_{\cdot i\cdot} - \bar{X})^2$	$l - 1$
Means of columns	$kl \sum_{j=1}^{m} (\bar{X}_{\cdot\cdot j} - \bar{X})^2$	$m - 1$
Interaction between rows and groups.	$m \sum_{h=1}^{k} \sum_{i=1}^{l} (\bar{X}_{hi\cdot} - \bar{X}_{h\cdot\cdot} - \bar{X}_{\cdot i\cdot} + \bar{X})^2$	$(k - 1)(l - 1)$
Interaction between rows and columns . .	$l \sum_{h=1}^{k} \sum_{j=1}^{m} (\bar{X}_{h\cdot j} - \bar{X}_{h\cdot\cdot} - \bar{X}_{\cdot\cdot j} + \bar{X})^2$	$(k - 1)(m - 1)$
Interaction between columns and groups . .	$k \sum_{i=1}^{l} \sum_{j=1}^{m} (\bar{X}_{\cdot ij} - \bar{X}_{\cdot i\cdot} - \bar{X}_{\cdot\cdot j} + \bar{X})^2$	$(l - 1)(m - 1)$
Interaction among rows, columns, and groups .	$\sum_{h=1}^{k} \sum_{i=1}^{l} \sum_{j=1}^{m} (X_{hij} - \bar{X}_{hi\cdot} - \bar{X}_{h\cdot j} - \bar{X}_{\cdot ij}$ $+ \bar{X}_{h\cdot\cdot} + \bar{X}_{\cdot i\cdot} + \bar{X}_{\cdot\cdot j} - \bar{X})^2$	$(k - 1)(l - 1)(m - 1)$
Total	$\sum_{h=1}^{k} \sum_{i=1}^{l} \sum_{j=1}^{m} (X_{hij} - \bar{X})^2$	$klm - 1$

TABLE 42

First Flowering Date (Day of Year) of 5 Varieties of Plants at 6 Stations Over a Period of 3 Years

	Hazel	Colts-foot	Anem-one	Black-thorn	Mustard	Total
Broadchalke—						
1932	57	67	95	102	123	444
1933	46	72	90	88	101	397
1934	28	66	89	109	113	405
Total	131	205	274	299	377	1246
Bratton—						
1932	26	44	92	96	93	351
1933	38	68	89	89	110	394
1934	20	64	106	106	115	400
Total	84	176	291	291	318	1145
Lenham—						
1932	48	61	78	99	113	339
1933	35	60	89	87	109	380
1934	48	75	95	113	111	442
Total	131	196	262	299	333	1221
Dorstone—						
1932	50	68	85	117	124	444
1933	37	65	74	93	102	371
1934	19	61	80	107	118	385
Total	106	194	239	317	344	1200
Coaley—						
1932	23	74	105	103	120	425
1933	36	47	85	90	101	359
1934	18	69	85	105	111	388
Total	77	190	275	298	332	1172
Ipswich—						
1932	39	57	91	102	112	401
1933	39	61	82	93	104	379
1934	43	61	98	98	112	412
Total	121	179	271	293	328	1192
All stations—						
1932	243	371	546	619	685	2464
1933	231	373	509	540	627	2280
1934	176	396	542	638	680	2432
Total	650	1140	1597	1797	1992	7176

the grand total divided by the number of items (5 × 6 × 3 = 90). This can readily be done on a calculating machine. We find

$$(57)^2 + (67)^2 + \cdots + (112)^2 - \frac{(7176)^2}{90}$$

$$= 644{,}074 - \frac{51{,}494{,}976}{90} = 644{,}074 - 572{,}166.4 = 71{,}907\ 6$$

For plants and stations we have Table 42A. From this table

TABLE 42A

Plants and Stations

	Hazel	Colts-foot	Anem-one	Black-thorn	Mustard	Total
Broadchalke	131	205	274	299	337	1246
Bratton . .	84	176	276	291	318	1145
Lenham	131	196	262	299	333	1221
Dorstone	106	194	239	317	344	1200
Coaley	777	190	275	298	332	1172
Ipswich . .	121	179	271	293	328	1192
Total .	650	1140	1597	1797	1992	7176

we calculate as we did for the double classification table of section 64,

$$\tfrac{1}{3}[(131)^2 + (205)^2 + \cdots + (328)^2] - \tfrac{1}{90}\,(7176)^2$$

$$= \tfrac{1}{3} \times 1{,}916{,}812 - 572{,}166.4 = 66{,}770.9\dot{3}$$

The reason that we take one-third of the sum of squares is that each value such as 131 is the total of three items (flowering dates for 1932, 1933, 1934). The value $66{,}770.9\dot{3}$ is the subtotal sum of squares of deviations for plants and stations.

The sum of squares of deviations for plants is

$$\tfrac{1}{18}\,[(650)^2 + (1140)^2 + \cdots + (1992)^2] - \tfrac{1}{90}\,(7176)^2$$

$$= \tfrac{1}{18} \times 11{,}469{,}782 - 572{,}166.4 = 65{,}04\dot{3}.7\dot{1}$$

Here each value such as 650 is the total of 18 items (flowering dates for 3 years at 6 stations), and we must divide the sum of squares by 18.

The sum of squares of deviations for stations is

$$\tfrac{1}{15}\,[1246)^2 + (1145)^2 + \cdots + (1192)^2] - \tfrac{1}{90}\,(7176)^2$$

$$= \tfrac{1}{15} \times 8{,}588{,}830 - 572{,}166.4 = 422\ 2\dot{6}$$

The interaction term for plants and stations is the remainder,

$$66{,}770.9\dot{3} - (65{,}043.7\dot{1} + 422\ 2\dot{6}) = 1304.9\dot{5}$$

For plants and years we have Table 42B.

TABLE 42B

PLANTS AND YEARS

	Hazel	Coltsfoot	Anemone	Black-thorn	Mustard	Total
1932	243	371	546	619	685	2464
1933	231	373	509	540	627	2280
1934	176	396	542	628	680	2432
Total .	650	1140	1597	1797	1992	7176

The subtotal sum of squares of deviations is

$$\tfrac{1}{6}[(243)^2 + (371)^2 + \cdots + (680)^2] - \tfrac{1}{90}\,(7176)^2$$

$$= \tfrac{1}{6} \times 3{,}834{,}492 - 572{,}166.4 = 66{,}915.6$$

The sum of squares of deviations for years is

$$\tfrac{1}{30}[(2464)^2 + (2280)^2 + (2432)^2] - \tfrac{1}{90}(7176)^2$$

$$= \tfrac{1}{30} \times 17{,}184{,}320 - 572{,}166.4 = 644.2\dot{6}$$

The sum of squares of deviations for plants has already been found to be $65{,}043.7\dot{1}$. Consequently the interaction term for plants and years is

$$66{,}915.6 - (644.2\dot{6} + 65{,}043.7\dot{1}) = 1227.2\dot{2}$$

The table for stations and years is Table 42C.

TABLE 42C

STATIONS AND YEARS

	Broad-chalke	Bratton	Lenham	Dorstone	Coaley	Ipswich	Total
1932	444	351	399	444	425	401	2464
1933	397	394	380	371	359	379	2280
1934	405	400	442	385	388	412	2432
Total	1246	1145	1221	1200	1172	1192	7176

Since we have the sums of squares of deviations for stations and for years ($422.2\dot{6}$ and $644.2\dot{6}$ respectively) we need only to find the subtotal sum of squares of deviations. This is

$$\tfrac{1}{5}[(444)^2 + (351)^2 + \cdots + (412)^2] - \tfrac{1}{90}(7176)^2$$

$$= \tfrac{1}{5} \times 8{,}588{,}830 - 572{,}166.4 = 2515.6$$

The interaction between stations and years is

$$2515\ 6 - (422.2\dot{6} + 644.2\dot{6}) = 1449.0\dot{6}$$

We can now find the sum of squares of deviations for the interaction among plants, stations, and years by subtracting from the total the terms for plants, stations, and years, and the single interaction terms, plants × stations, plants × years, and stations × years. This gives

$$71{,}907.4 - (65{,}043.7\dot{1} + 422.2\dot{6} + 644.2\dot{6} + 1304.9\dot{5}$$

$$+ 1227.2\dot{2} + 1449.0\dot{6}) = 1816.1\dot{1}$$

We are now ready to form the analysis of variance table (Table 42D).

TABLE 42D

Analysis of Variance

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Plants	65,043 7$\dot{1}$	5−1= 4	16,260 92$\dot{7}$
Stations ...	422 2$\dot{6}$	6−1= 5	84 453
Years . .	644 2$\dot{6}$	3−1= 2	322 133
Plants × stations	1,304 95	4×5=20	65 247
Plants × years	1,227.2$\dot{2}$	4×2= 8	153 402$\dot{7}$
Stations × years	1,449 06	5×2=10	144 90$\dot{6}$
Plants × stations × years	1,816 1$\dot{1}$	4×5×2=40	45 4027
Total	71,907 6	5×6×3−1=89	

If we wish to test any variance, that is, any mean square deviation, we use the interaction of plants, stations, and years for the comparison term. For example, if we wish to test the variation from year to year we take

$$z = \tfrac{1}{2} \log_e \frac{322\ 13}{45.403} = 0\ 9797, \quad n_1 = 2, \quad n_2 = 40$$

This value is outside the 5 per cent point and is very close to the 1 per cent point.

In double or multiple classification the case of unequal frequencies in the various classes presents some difficulties. Those interested in this case are referred to the original papers in which it has been treated.*

65. Analysis of covariance.† The covariance between N pairs of values of X and Y has been defined (see section 18) as $\Sigma xy/N$, where $x = X - \overline{X}$, $y = Y - \overline{Y}$. It is quite possible to

* See F Yates, "The analysis of multiple classifications with unequal numbers in the different classes," *Journal of the American Statistical Association*, vol. 29, 1934, pp 51–66; and the references contained therein

† For a fuller treatment of this topic consult the following references:

R. A. Fisher, "Statistical Methods for Research Workers," section 49.1.

J. Wishart and H G Sanders, "Principles and Practice of Field Experimentation," Empire Cotton Growing Corporation, London, 1936

J. Wishart, "Tests of significance in analysis of covariance," *Supplement to the Journal of the Royal Statistical Society*, vol. 3, 1936, pp. 79–82, and further references contained therein.

analyze the sum of products into components just as we analyzed the sum of squares. In this way we are able to obtain estimates of the covariance, and also of the regression and correlation coefficients, which are freed from class or other effects. The analysis of covariance has been used in agricultural and biological problems by Fisher and others. Bailey * has given a number of miscellaneous examples of its use and has emphasized its applicability to time series in economics.

Suppose that we have a table (Table 43) similar to Table 31

TABLE 43

Class	1	···	k
	(X_{11}, Y_{11})	···	(X_{1k}, Y_{1k})
	. . .	. .	. . .
	(X_{m1}, Y_{m1})	···	(X_{mk}, Y_{mk})
Mean	$(\bar{X}_1, \bar{Y}_1)$	··	$(\bar{X}_k, \bar{Y}_k)$

but with two variables instead of one. That is, we have k classes with m individuals in each. With the ith individual in the jth class is associated a pair of values X_{ij}, Y_{ij}, which may be regarded as the measures of a certain pair of attributes or characteristics. If $\bar{X}_j$, $\bar{Y}_j$ are the means of X and Y respectively in the jth class, then the sum of products Σxy can be written

$$\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\bar{X})(Y_{ij}-\bar{Y})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{m}[(X_{ij}-\bar{X}_j)+(\bar{X}_j-\bar{X})][(Y_{ij}-\bar{Y}_j)+(\bar{Y}_j-\bar{Y})]$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{m}(X_{ij}-\bar{X}_j)(Y_{ij}-\bar{Y}_j)+\sum_{j=1}^{k}\sum_{i=1}^{m}(\bar{X}_j-\bar{X})(Y_{ij}-\bar{Y}_j)$$

$$+\sum_{j=1}^{k}\sum_{i=1}^{m}(\bar{Y}_j-\bar{Y})(X_{ij}-\bar{X}_j)+\sum_{j=1}^{k}\sum_{i=1}^{m}(\bar{X}_j-\bar{X})(\bar{Y}_j-\bar{Y})$$

* A. L Bailey, "The analysis of covariance," *Journal of the American Statistical Association*, vol. 26, 1931, pp 424–435.

$$= \sum_{j=1}^{k} \sum_{i=1}^{m} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)$$

$$+ m \sum_{j=1}^{k} (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y}) \tag{43}$$

The term $\Sigma_j\Sigma_i(\bar{X}_j - \bar{X})(Y_{ij} - \bar{Y}_j)$ vanishes, since it is equal to

$$\sum_{j=1}^{k} (\bar{X}_j - \bar{X}) \sum_{i=1}^{m} (Y_{ij} - \bar{Y}_j) = \sum_{j=1}^{k} (\bar{X}_j - \bar{X}) m(\bar{Y}_j - \bar{Y}_j) = 0$$

Similarly the corresponding term with X and Y interchanged vanishes. The fundamental identity (43) expresses the fact that the total sum of products of deviations is equal to the sum of products of deviations within classes plus the sum of products of the deviations of the class means from the general means, multiplied by the number of individuals in each class.

The sum of products can be broken up into more components just as can the sum of squares It is not necessary to illustrate further subdivision, however, as it can be effected in a manner entirely analogous to the subdivision in the analysis of variance.

As an example of the analysis of covariance, consider the following data on the yield of wheat in bushels per acre and the production cost per bushel for five farms in each of three districts.

TABLE 44

District I		District II		District III	
Yield (bushels per acre)	Cost per bushel	Yield (bushels per acre)	Cost per bushel	Yield (bushels per acre)	Cost per bushel
9	$1 80	18	$1 00	14	$1 00
11	1 40	18	0 50	10	1 50
8	2 00	20	0 70	13	0 80
9	1 50	16	0 70	15	0 70
11	1 70	9	2 00	7	2 10

Designating yield by X and cost by Y, we form Table 45, in which are also listed X^2, XY, Y^2.

TABLE 45

	X	Y	X^2	XY	Y^2
District I	9	1 80	81	16 2	3.24
	11	1 40	121	15 4	1 96
	8	2 00	64	16 0	4 00
	9	1 50	81	13 5	2 25
	11	1 70	121	18 7	2 89
Subtotal	48	8 40	468	79 8	14 34
District II	18	1 00	324	18 0	1 00
	18	0 50	324	9 0	0 25
	20	0 70	400	14 0	0 49
	16	0 70	256	11 2	0 49
	9	2 00	81	18 0	4 00
Subtotal	81	4 90	1385	70 2	6 23
District III	14	1 00	196	14 0	1 00
	10	1 50	100	15 0	2 25
	13	0 80	169	10 4	0 64
	15	0 70	225	10 5	0 49
	7	2 10	49	14 7	4 41
Subtotal	59	6 10	739	64 6	8 79
Total. .	188	19 40	2592	214 6	29 36

TABLE 46

AMONG DISTRICTS

District	X	Y	X^2	XY	Y^2
I	48	8 40	2,304	403 2	70 56
II	81	4 90	6,561	396 9	24 01
III	59	6 10	3,481	359 9	37 21
Total	188	19 40	12,346	1160 0	131 78

In computing, from Tables 45 and 46, the sums of products of deviations and the sums of squares of deviations, we make use of the formulas

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}, \quad \Sigma xy = \Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N},$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$$

For the total we have

$$\Sigma x^2 = 2592 - (188)^2/15 = 2592 - 2356\ 2\dot{6} = 235.7\dot{3}$$

$$\Sigma xy = 214.\dot{6} - 188 \times 19.40/15 = 214.6 - 243.14\dot{6} = -\ 28.54\dot{6}$$

$$\Sigma y^2 = 29.36 - (19.4)^2/15 = 29.36 - 25.090\dot{6} = 4.269\dot{3}$$

Among districts,

$$\Sigma x^2 = \tfrac{1}{5}\times 12346 - \tfrac{1}{15}(188)^2 = 2469\ 2 - 2356\ 2\dot{6} = 112.9\dot{3}$$

$$\Sigma xy = \tfrac{1}{5}\times 1160.0 - \tfrac{1}{15}\times 188 \times 19.4 = 232 - 243.14\dot{6} = -11.14\dot{6}$$

$$\Sigma y^2 = \tfrac{1}{5}\times 131.78 - \tfrac{1}{15}(19.4)^2 = 26.356 - 25.090\dot{6} = 1.265\dot{3}$$

Within districts,

I. $\Sigma x^2 = 468 - (48)^2/5 = 468 - 460.8 = 7.2$

$\Sigma xy = 79.8 - 48 \times 8.4/5 = -\ 0.84$

$\Sigma y^2 = 14.34 - (8.4)^2/5 = 14.34 - 14.112 = 0.228$

II. $\Sigma x^2 = 1385 - (81)^2/5 = 1385 - 1312.2 = 72.8$

$\Sigma xy = 70.2 - 81 \times 4.9/5 = 70\ 2 - 79.38 = -\ 9.18$

$\Sigma y^2 = 6.23 - (4.9)^2/5 = 6.23 - 4.802 = 1.428$

III. $\Sigma x^2 = 739 - (59)^2/5 = 739 - 696.2 = 42.8$

$\Sigma xy = 64.6 - 59 \times 6.1/5 = 64.6 - 71.98 = -\ 7.38$

$\Sigma y^2 = 8.79 - (6.1)^2/5 = 8.79 - 7.442 = 1.348$

Compiling these results, we have Table 47. From this table we can calculate the regression coefficients for the different components. Within districts we find

$$b_1 = -\frac{17.40}{122.8} = -\ 0.1417$$

while among districts

$$b_2 = -\frac{11.14\dot{6}}{112.9\dot{3}} = -\ 0.0987$$

TABLE 47

ANALYSIS OF VARIANCE AND COVARIANCE

	Σx^2	Σxy	Σy^2	Degrees of freedom
District I	7 2	−0 84	0 228	4
II	72 8	−9 18	1 428	4
III	42 8	−7 38	1 348	4
Within districts	122 8	−17 40	3 004	12
Among districts	112 9$\dot{3}$	−11 14$\dot{6}$	1 265$\dot{3}$	2
Total .	235 7$\dot{3}$	−28 54$\dot{6}$	4 269$\dot{3}$	14

Let us first test whether the coefficient of regression within districts is significant. We may use the t test of section 43, or may use the analysis of variance method Let us use the latter. For the sum of squares of deviations from the regression line, $y' = b_1x$, we have

$$\Sigma(y - b_1x)^2 = \Sigma y^2 - 2b_1\Sigma xy + b_1^2\Sigma x^2$$

which, since $b_1 = \Sigma xy/\Sigma x^2$, reduces to

$$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} = 3.004 - \frac{(-17.40)^2}{122.8} = 3.004 - 2.46547 = 0.53853$$

The analysis is summarized in Table 48.

TABLE 48

Analysis of Variance Within Districts

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Residuals	0 53853	11	0 0490
Regression	2 46547	1	2 4655
Total	3 004	12	

We find $z = \frac{1}{2} \log_e (2.4655/0.0490) = 1.959$, which for $n_1 = 1$ and $n_2 = 11$ is highly significant.

Next we analyze the residual variance. For the regression among districts we have, as above,

$$\Sigma(y - b_2x)^2 = 1.265\dot{3} - \frac{(-11.14\dot{6})^2}{112.9\dot{3}}$$

$$= 1.265\dot{3} - 1.10019 = 0.16514$$

The sum of squares of differences between regressions is *

$$\frac{(-17\ 40)^2}{122.8} + \frac{(-11\ 14\dot{6})^2}{112.9\dot{3}} - \frac{(-28\ 54\dot{6})^2}{235.73}$$

$$= 2.46547 + 1.10019 - 3.45692 = 0.10874$$

The sum of squares of the deviations among districts is

$$0.16514 + 0.10874 = 0.27388$$

The sum of squares of deviations for the "total" is

$$4.269\dot{3} - \frac{(-28.54\dot{6})^2}{235.7\dot{3}} = 4.269\dot{3} - 3.45692 = 0.81241$$

and the corresponding sum within districts has already been found to be 0.53853.

These results are recapitulated in Table 49. Any desired tests can be made by computing the appropriate values of z, using the

* See Wishart and Sanders, *op. cit.*

mean square deviation within districts as the error mean square.

TABLE 49

Analysis of Residual Variance

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Between regressions	0 10874	1	0 10874
Residuals from the "among" regression	0 16514	1	0 16514
Among districts	0 27388	2	0 13694
Within districts	0 53853	11	0 04896
Total	0 81241	13	

To be noted is the fact that the estimated standard deviation has been decreased from $(3\ 004/12)^{1/2} = 0.503$, to $(0.04896)^{1/2} = 0.221$; that is, it has been cut to less than half, thus greatly increasing the accuracy of any tests made.

EXERCISES

1. (*a*) From the data of Table O, page 63, can it be concluded that either sex has a greater variability in number of red blood cells than the other? (*b*) Can it be concluded that either sex has a greater variability in amount of hemoglobin?

2. Analyze the variance for the linear regression found in exercise 2, page 43, and make a test of significance

3. Make an analysis of variance for the parabolic regression obtained in exercise 5, page 45, and make tests of significance

4. Analyze the variance for the multiple regression obtained in exercise 8, page 45, and make a test of significance.

5. (*a*) Analyze the variance for the correlation ratio found in exercise 3, page 62. (*b*) Test the significance of this correlation ratio (*c*) Test the linearity of the regression

6. Make an analysis of variance for the correlation ratio found in exercise 4, page 64. Test (*a*) the significance of the correlation ratio, and (*b*) the linearity of the regression

7. Analyze the variance in Table W into that within batches and that among batches, and test for significance.

TABLE W

Breaking Strength (Pounds Tension) of 10 Batches of Cement Briquettes

1	2	3	4	5	6	7	8	9	10
518	508	554	555	536	544	578	530	590	542
560	574	598	567	492	502	532	564	554	556
538	528	579	550	528	548	562	536	530	590
510	534	538	535	572	562	524	540	572	546
544	538	544	540	506	534	548	530	525	522

8. Table Y gives porosity readings on 3 lots of condenser paper There are 3 readings on each of 9 rolls from each lot Determine whether there are significant variations (*a*) among readings within rolls, (*b*) among rolls within lots, and (*c*) among lots (Western Electric Co data)

TABLE Y

Porosity Readings on Condenser Paper

Lot number	Reading number	Roll number								
		1	2	3	4	5	6	7	8	9
I	1	1 5	1 5	2 7	3 0	3 4	2 1	2 0	3 0	5 1
	2	1 7	1 6	1 9	2 4	5 6	4 1	2 5	2 0	5 0
	3	1 6	1 7	2 0	2 6	5 6	4 6	2 8	1 9	4 0
II	1	1 9	2 3	1 8	1 9	2 0	3 0	2 4	1 7	2 6
	2	1 5	2 4	2 9	3 5	1 9	2 6	2 0	1 5	4 3
	3	2 1	2 4	4 7	2 8	2.1	3 5	2 1	2 0	2 4
III	1	2 5	3 2	1 4	7 8	3 2	1 9	2 0	1 1	2 1
	2	2 9	5 5	1 5	5 2	2 5	2 2	2 4	1 4	2 5
	3	3 3	7 1	3 4	5 0	4 0	3 1	3 7	4 1	1 9

9. Table Z gives impact strength readings, in foot-pounds, on 5 lots of insulating material One specimen from each of 20 different sheets was tested from each lot The first 10 specimens were cut along the lengthwise direction of the sheets, the second 10 specimens were cut along the crosswise direction. Determine whether there are significant variations (*a*) among lots, and (*b*) between lengthwise and crosswise specimens (Western Electric Co data)

TABLE Z

Impact Strength Readings (Foot-Pounds) on 5 Lots of Insulating Material

	Specimen number	Lot number				
		I	II	III	IV	V
Lengthwise specimens	1	1 15	1 16	0 79	0 96	0 49
	2	0 84	0 85	68	82	61
	3	88	1 00	64	98	59
	4	91	1 08	72	93	51
	5	86	0 80	63	.81	53
	6	88	1 01	59	79	72
	7	92	1 14	81	79	67
	8	87	0 87	65	86	47
	9	93	0 97	64	84	44
	10	95	1 09	.75	92	48
Crosswise specimens	11	0 89	0 86	0 52	0 86	0 52
	12	69	1 17	52	1 06	53
	13	46	1 18	80	0 81	47
	14	85	1 32	.64	97	47
	15	73	1 03	63	90	.57
	16	67	0 84	58	93	54
	17	78	0 89	.65	87	56
	18	77	0 84	60	88	55
	19	80	1 03	71	89	45
	20	79	1 06	59	82	60

10. The followıng data are the thıcknesses of coatıng, in 0 0001 ın , on fiber strıps sprayed wıth varnısh Measurements were taken at 5 dıfferent points on each of the 3 strıps selected from each of 5 lots Test for sıgnıficance the varıance (*a*) among poınts wıthın strips, (*b*) among strıps wıthın lots, and (*c*) among lots (Western Electrıc Co. data)

Lot I

Strıp number	Poınt number				
	1	2	3	4	5
1	10	8	10	9	7
2	8	8	8	8	10
3	8	10	10	6	7

Lot II

Strıp number	Poınt number				
	1	2	3	4	5
1	13	12	12	12	13
2	10	9	13	11	8
3	11	8	10	12	12

Lot III

Strıp number	Poınt number				
	1	2	3	4	5
1	12	13	14	17	16
2	17	10	13	10	14
3	12	11	13	16	13

Lot IV

Strıp number	Poınt number				
	1	2	3	4	5
1	14	13	17	11	11
2	11	9	13	11	12
3	17	13	14	13	8

Lot V

Strıp number	Poınt number				
	1	2	3	4	5
1	9	13	17	13	11
2	8	11	10	12	11
3	7	14	14	9	9

11. Table AA gives the initial weights and the average daily gains of 4 lots of 10 pigs each Each lot was fed on a different diet Make an analysis of variance and covariance, and test the significance of the differences among the adjusted or residual mean gains of the different lots

TABLE AA

INITIAL WEIGHTS (X POUNDS) AND AVERAGE DAILY GAINS (Y POUNDS PER DAY) OF 4 LOTS OF PIGS

Lot 1		Lot 2		Lot 3		Lot 4	
Initial weight X	Daily gain Y	Initial weight X	Daily gain Y	Initial weight X	Daily gain Y	Initial weight X	Daily gain Y
36	1 33	38	1 25	45	1 22	38	1 35
65	1 13	60	1 39	59	1 79	73	1 60
44	1 80	41	1 57	38	1 31	40	1 26
51	1 48	50	1 29	53	1 50	43	1 15
66	1 76	61	1 25	50	1 31	44	1 54
44	1 42	44	1 20	45	1 55	48	1 24
57	1 90	60	1 40	56	1 70	50	1 47
79	1 67	71	1 29	61	1 40	62	1 29
41	1 31	53	1 30	39	1 36	51	1 41
57	1 34	54	1 46	59	1 53	58	1 35

CHAPTER IX

EXPERIMENTAL DESIGN

66. Randomized blocks. The realization is becoming stronger that, in order to yield the best results and the greatest possible information, an experiment should be properly planned before it is performed. The development of the analysis of variance and the improvement of experimental design have proceeded simultaneously. In agricultural science particularly, it has been found desirable to design experiments so that the analysis of variance can be conveniently and correctly applied.

One field arrangement that has been found extremely useful, and at the same time specially suited to the application of the analysis of variance, is that of *randomized blocks*. Consider, for simplicity, an experiment which is to be made on three varieties of wheat to ascertain which has the greatest yield in bushels per acre. We might take four blocks of land, divide each block into three strips, and sow each strip of a given block with a different variety, arranging them at random. Such an experiment would be represented by the following diagram, in which v_1, v_2, v_3 are the three varieties, and the numbers shown in the various strips are the yields in bushels per acre.

1	2	3	4
v_3 9	v_1 10	v_3 20	v_2 14
v_2 11	v_3 18	v_1 16	v_1 12
v_1 8	v_2 18	v_2 9	v_3 7

When the data are arranged in tabular form we have Table 50.

TABLE 50

YIELDS IN BUSHELS PER ACRE OF 3 VARIETIES OF WHEAT

Variety	Block				Total
	1	2	3	4	
1	8	10	16	12	46
2	11	18	9	14	52
3	9	18	20	7	54
Total	28	46	45	33	152

They can be analyzed as before by formulas (40), (41), and (42) of section 64.

The total sum of squares of deviations is

$$\Sigma_j\Sigma_i(X_{ij} - \bar{X})^2 = \Sigma_j\Sigma_i X_{ij}^2 - \frac{(\Sigma_j\Sigma_i X_{ij})^2}{N}$$

$$= (8)^2 + (10)^2 + (16)^2 + (12)^2 + (11)^2 + (18)^2 + (9)^2$$

$$+ (14)^2 + (9)^2 + (18)^2 + (20)^2 + (7)^2 - \frac{(152)^2}{12}$$

$$= 2140 - 1925.\dot{3} = 214.\dot{6}$$

The sum of squares of block differences (due to differences in soil fertility, etc) is

$$m\Sigma_j(\bar{X}_{\cdot j} - \bar{X})^2 = \frac{1}{m}\Sigma_j(\Sigma_i X_{ij})^2 - \frac{1}{N}(\Sigma_j\Sigma_i X_{ij})^2$$

$$= \tfrac{1}{3}[(28)^2 + (46)^2 + (45)^2 + (33)^2] - \tfrac{1}{12}(152)^2$$

$$= 2004.\dot{6} - 1925.\dot{3} = 79\ \dot{3}$$

For variety differences we find

$$k\Sigma_i(\bar{X}_{i\cdot} - \bar{X})^2 = \frac{1}{k}\Sigma_i(\Sigma_j X_{ij})^2 - \frac{l}{N}(\Sigma_j\Sigma_i X_{ij})^2$$

$$= \tfrac{1}{4}[(46)^2 + (52)^2 + (54)^2] - \tfrac{1}{12}(152)^2$$

$$= 1934 - 1925.\dot{3} = 8.\dot{6}$$

The error term is

$$214\ \dot{6} - 79.\dot{3} - 8.\dot{6} = 126.\dot{6}$$

so that the analysis of variance table is as follows:

TABLE 51

	Sum of squares of deviations	Degrees of freedom	Mean square deviation
Blocks .	79 $\dot{3}$	$4 - 1 = 3$	26.$\dot{4}$
Varieties.	8 $\dot{6}$	$3 - 1 = 2$	4 3
Error .	126 6	$3 \times 2 = 6$	21 $\dot{1}$
Total .	214 6	$12 - 1 = 11$	

The mean square deviation for varieties is less than that for error.

A similar analysis could be made if different treatments were applied to the same variety.

If each variety in the foregoing illustration were treated with two different kinds of fertilizer we should have six different combinations

$$v_1t_1,\quad v_1t_2,\quad v_2t_1,\quad v_2t_2,\quad v_3t_1,\quad v_3t_2$$

where of course v refers to variety and t to treatment. Each block would be divided into six different plots and a typical block would be that shown in the diagram. That is, each block would be divided into three strips, to which would be assigned at random the three varieties. Each strip would be subdivided into two plots, and the treatments would be allocated to them at random. The results could be tabulated as in Table 52.

$v_2\ t_1$	$v_2\ t_2$
$v_1\ t_2$	$v_1\ t_1$
$v_3\ t_3$	$v_3\ t_1$

These data could be analyzed in precisely the same way as were the phenological data (first flowering dates of plants).

TABLE 52

YIELDS IN BUSHELS PER ACRE OF THREE VARIETIES OF WHEAT TREATED WITH TWO DIFFERENT FERTILIZERS

	Block				Total
	1	2	3	4	
Variety 1					
Treatment 1	8	10	16	12	46
Treatment 2.	7	8	12	13	40
Total .	15	18	28	25	86
Variety 2					
Treatment 1.	11	18	9	14	52
Treatment 2	10	16	10	12	48
Total .	21	34	19	26	100
Variety 3					
Treatment 1	9	18	20	7	54
Treatment 2	10	20	20	10	60
Total .	19	38	40	17	114
All varieties					
Treatment 1	28	46	45	33	152
Treatment 2 .	27	44	42	35	148
Total. .	55	90	87	68	300

It should be noted that when we have only two subdivisions, as for treatments in the foregoing illustration, the sum of squares of deviations can sometimes be calculated more simply as follows: The totals 152 and 148 represent 12 plots each, and according to the method already employed the sum of squares would be

$$\tfrac{1}{12}[(152)^2 + (148)^2] - \tfrac{1}{24}(300)^2 = 0\ \dot{6}$$

However, the difference between 152 and 148 represents 24 plots, and we obtain the same result by taking

$$\tfrac{1}{24}(152 - 148)^2 = 0\ \dot{6}$$

In general, the sum of squares of deviations of two quantities X_1 and X_2 from their mean is

$$(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = \frac{(X_1 - X_2)^2}{2} \tag{1}$$

Their variance is $(X_1 - X_2)^2/4$ and their standard deviation $X_1 - X_2\,|/2$.

67. Latin square. One arrangement frequently used is the *Latin square.* If we are testing m varieties (or treatments), a block of land is divided into a checkerboard arrangement of m rows and m columns, and the varieties are distributed at random in the plots, with the restriction that each variety occurs once and but once in each row and also in each column. If A, B, C, D, E are five varieties, we can form a five-by-five Latin square, a typical example of which is shown in Fig. 12. It will be noted that each letter occurs once and only once in each row and each column. The fundamental identity for the Latin square of order m (m rows, m columns, m varieties) is

C	*A*	*E*	*D*	*B*
D	*B*	*A*	*E*	*C*
A	*D*	*C*	*B*	*E*
B	*E*	*D*	*C*	*A*
E	*C*	*B*	*A*	*D*

FIG 12.—Latin Square.

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 = m\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X})^2 + m\sum_{j=1}^{m}(\bar{X}_{\cdot j} - \bar{X})^2$$

$$+ m\sum_{k=1}^{m}(\bar{X}_k - \bar{X})^2 + \sum_{i=1}^{m}\sum_{j=1}^{m}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} - \bar{X}_k + 2\bar{X})^2 \tag{2}$$

In this formula i refers to row, j to column, and k to variety[*] (or treatment). That is, X_{ij} is the item in the ith row and jth column, $\bar{X}_{i\cdot}$ is the mean of the ith row, $\bar{X}_{\cdot j}$ the mean of the jth column, $\bar{X}_k$ the mean of the kth variety, and $\bar{X}$ the general mean The analysis of variance table showing the degrees of freedom is given below (Table 53).

TABLE 53

ANALYSIS OF VARIANCE FOR A LATIN SQUARE OF ORDER m

	Sum of squares of deviations	Degrees of freedom
Rows	$m\Sigma_i(\bar{X}_{i\cdot} - \bar{X})^2$	$m - 1$
Columns	$m\Sigma_j(\bar{X}_{\cdot j} - \bar{X})^2$	$m - 1$
Varieties	$m\Sigma_k(\bar{X}_k - \bar{X})^2$	$m - 1$
Error	$\Sigma_i\Sigma_j(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} - \bar{X}_k + 2\bar{X})^2$	$(m - 1)(m - 2)$
Total	$\Sigma_i\Sigma_j(X_{ij} - \bar{X})^2$	$m^2 - 1$

The variation in the totals of rows and of columns gives an indication of the amount of soil heterogeneity running in two directions at right angles to each other, and it is the object of the Latin square arrangement to remove the effect of this heterogeneity. For a square larger than eight-by-eight, the rows and columns tend to become too long, and the efficiency of the design is impaired. For a four-by-four square there are three degrees of freedom for varieties and six for error, and we find, by reference to Snedecor's tables, that the mean square deviation for varieties can be judged significant at the 5 per cent level if it is 4.76 times that for error. But for a three-by-three square there are only two degrees of freedom for varieties and the same number for error, so that the variety mean square will have to be nineteen times the error mean square before it can be judged significant. Consequently, squares larger than eight-by-eight or smaller than four-by-four are not recommended in practice.

For the sake of simplicity, however, we shall give the analysis

[*] The subscript k would refer to the variety occurring in row i and column j.

of a three-by-three square. In this square (Fig. 13) the v's in the compartments refer to varieties of wheat; the numbers in the compartments are the corresponding yields in bushels per acre.

				Total
	v_2 11	v_3 9	v_1 8	28
	v_1 10	v_2 18	v_3 18	46
	v_3 20	v_1 16	v_2 9	45
Total	41	43	35	119

	Varieties		
	v_1	v_2	v_3
	8	11	9
	10	18	18
	16	9	20
Total	34	38	47

Fig. 13.—Latin Square Showing Yields of Three Varieties of Wheat.

The calculation of the sum of squares of deviations is as follows:

Total

$$(11)^2 + (9)^2 + (8)^2 + (10)^2 + (18)^2 + (18)^2 + (20)^2 + (16)^2 + (9)^2 - \frac{(119)^2}{9} = 1751 - 1573.\dot{4} = 177.\dot{5}$$

Rows

$$\tfrac{1}{3}[(28)^2 + (46)^2 + (45)^2] - \tfrac{1}{9}(119)^2 = 1641.\dot{6} - 1573.\dot{4} = 68.\dot{2}$$

Columns

$$\tfrac{1}{3}[(41)^2 + (43)^2 + (35)^2] - \tfrac{1}{9}(119)^2 = 1585 - 1573.\dot{4} = 11.\dot{5}$$

Varieties

$$\tfrac{1}{3}[(34)^2 + (38)^2 + (47)^2] - \tfrac{1}{9}(119)^2 = 1603 - 1573.\dot{4} = 29.\dot{5}$$

Error

$$177.\dot{5} - (68.\dot{2} + 11.\dot{5} + 29.\dot{5}) = 177.\dot{5} - 109.\dot{3} = 68.\dot{2}$$

We have calculated the error term as a remainder, in which way it is usually found. For illustrative purposes we shall also calculate

$$\Sigma_i\Sigma_j(X_{ij} - \bar{X}_i - \bar{X}_{\cdot j} - \bar{X}_k + 2\bar{X})^2$$

$$= (11-\tfrac{28}{3}-\tfrac{41}{3}-\tfrac{38}{3}+2\times\tfrac{119}{9})^2 + (9-\tfrac{28}{3}-\tfrac{43}{3}-\tfrac{47}{3}+2\times\tfrac{119}{9})^2$$
$$+(8-\tfrac{28}{3}-\tfrac{35}{3}-\tfrac{34}{3}+2\times\tfrac{119}{9})^2+(10-\tfrac{46}{3}-\tfrac{41}{3}-\tfrac{34}{3}+2\times\tfrac{119}{9})^2$$
$$+(18-\tfrac{46}{3}-\tfrac{43}{3}-\tfrac{38}{3}+2\times\tfrac{119}{9})^2+(18-\tfrac{46}{3}-\tfrac{35}{3}-\tfrac{47}{3}+2\times\tfrac{119}{9})^2$$
$$+(20-\tfrac{45}{3}-\tfrac{41}{3}-\tfrac{47}{3}+2\times\tfrac{119}{9})^2+(16-\tfrac{45}{3}-\tfrac{43}{3}-\tfrac{34}{3}+2\times\tfrac{119}{9})^2$$
$$+9-\tfrac{45}{3}-\tfrac{35}{3}-\tfrac{38}{3}+2\times\tfrac{119}{9})^2$$

$$= \frac{1}{9^2}[(16)^2 + (-35)^2 + (19)^2 + (-35)^2$$
$$+ (19)^2 + (16)^2 + (19)^2 + (16)^2 + (-35)^2]$$
$$= \tfrac{5526}{81} = 68.\dot{2}$$

The analysis of variance table (Table 54) follows.

TABLE 54

Analysis of Variance Table for Latin Square of Wheat Yields

	Sum of squares of deviations	Degrees of freedom
Rows .	68 $\dot{2}$	$3 - 1 = 2$
Columns. . .	11 $\dot{5}$	$3 - 1 = 2$
Varieties. . .	29 $\dot{5}$	$3 - 1 = 2$
Error. . .	68 $\dot{2}$	$2 \times 1 = 2$
Total .	177 $\dot{5}$	$3^2 - 1 = 8$

68. Factorial design and orthogonality. In any sort of experiment it is usually better to vary several factors simultaneously rather than one at a time. For example, in an agricultural experiment on yields it is better to test several varieties with several different kinds of fertilizer, and even with varying degrees or levels of each kind of fertilizer. For if four different varieties were being tested with the same fertilizer we might find, for example, that the second variety gave the greatest yield. A separate investigation on this variety with three kinds of fertilizer might show that the first kind of fertilizer gave the best results. However, from these two experiments we could not tell but that the second kind of fertilizer, say, when applied to the fourth variety would give still better results. If the experiments were all combined into one, in which all four varieties are tested in conjunction with all three kinds of fertilizer, much more information will be elicited, since a large share of it is contained in the interactions among the various factors at work.

The method of experimentation in which two or more sets of factors, such as treatments and varieties, are taken in all combinations has been called *factorial design.** Factorial design had its inception in agriculture, and its greatest development has taken place in that science. However, it should doubtless find application in biology and medicine, and in testing materials and manufactured goods.

The usefulness of the analysis of variance in testing significance in factorial design consists in the multiplicity of ways in which the sum of squares of deviations can be split up. For example, consider the numbers X_1, X_2, X_3. Let us form the two linear expressions

$$X_1 - X_2$$

$$X_1 + X_2 - 2X_3$$

which, together with the sum

$$X_1 + X_2 + X_3$$

constitute a mutually *orthogonal* set. This means that the sum of products of corresponding coefficients in any two members of the set is zero, e.g., $1 \times 1 + (-1) \times 1 + 0 \times (-2) = 0$. The term "orthogonal" comes from geometry; we recall that, if a_1, b_1, c_1 and a_2, b_2, c_2 are the direction numbers of two lines, these lines will be orthogonal (that is, perpendicular) if $a_1a_2 + b_1b_2 + c_1c_2 = 0$. It is worth noting that, if a linear function of the X's is orthogonal to their sum, then the sum of the coefficients of this linear function

* See R. A. Fisher, "The Design of Experiments," Oliver and Boyd, Edinburgh and London; also F Yates, "Complex experiments," *Supplement to the Journal of the Royal Statistical Society*, vol. 2, 1935, pp 181–247.

is zero. Thus, for the above linear functions, we have $1 - 1 = 0$ and $1 + 1 - 2 = 0$.

When we have a set of linear combinations of a series of numbers, which, together with the sum of the numbers, constitutes a complete * mutually orthogonal set, we have one way of subdividing the sums of squares of deviations of these numbers from their mean. In the above example we have

$$\frac{(X_1 - X_2)^2}{1^2 + (-1)^2} + \frac{(X_1 + X_2 - 2X_3)^2}{1^2 + 1^2 + (-2)^2} = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2$$

as can readily be verified. The denominators on the left are the sums of squares of coefficients.

There are many ways of forming orthogonal sets. Thus, with four variables we might, to enumerate three such sets, have

$$\begin{array}{lcr}
X_1 - X_2 & X_1 + X_2 - X_3 - X_4 & -3X_1 - X_2 + X_3 + 3X_4 \\
X_1 + X_2 - 2X_3 & X_1 - X_2 & X_1 - X_2 - X_3 + X_4 \\
X_1 + X_2 + X_3 - 3X_4 & X_3 - X_4 & -X_1 + 3X_2 - 3X_3 + X_4
\end{array}$$

Each column, together with the sum, $X_1 + X_2 + X_3 + X_4$, constitutes a complete mutually orthogonal set. If the X's correspond to treatments, for example, any set may be regarded as comparisons, and, moreover, independent comparisons among them. (With four variables we must have three independent comparisons.) That is, if X_1 is the total yield of plots treated with the first fertilizer, and so on, then examining the first set we see that $X_1 - X_2$ compares the yield due to the first treatment with that due to the second, $X_1 + X_2 - 2X_3$ compares the sum of the yields due to the first and second treatments with twice that due to the third, while $X_1 + X_2 + X_3 - 3X_4$ compares the yields of the first three treatments with three times that of the fourth.

To see how this method of subdivision may be used to advantage in factorial design, let us consider an experiment in which we

* By a *complete* set we mean one that contains just as many linear combinations (including the sum of the numbers) as there are numbers.

are testing three kinds of fertilizer, *a*, *b*, *c*. We might not merely want to apply them alone, but in conjunction with one another, so that altogether we should have not three treatments, but eight, which might be designated by

$$abc, \quad ab, \quad ac, \quad bc, \quad a, \quad b, \quad c, \quad (1)$$

the symbol (1) denoting the absence of application of *a*, *b*, and *c*. The symbol *abc*, for example, may be regarded qualitatively as the treatment containing all three ingredients, or quantitatively as the total yield of plots treated with all three ingredients. We should have blocks of land with eight plots each, the treatments being scattered at random over the blocks, each block, however, containing all the treatments. A typical block is shown in the accompanying diagram. If we had five such blocks or *replications*, the degrees of freedom in the analysis of variance would be as follows:

ab	*b*	(1)	*abc*
c	*bc*	*ac*	*a*

	Degrees of freedom
Blocks..............	$5 - 1 = 4$
Treatments	$8 - 1 = 7$
Error..	$4 \times 7 = 28$
Total........... .	$5 \times 8 - 1 = 39$

The number of degrees of freedom for treatments is the number of independent comparisons.

The total or the mean yield of plots having the treatment *abc* could be calculated, similarly for *ab* and all the others, and we could determine whether the variation in these means was or was not accidental, and any one of them could be tested individually.

Instead of considering the treatments as they are, we might be more interested in subdividing them in another manner. For instance, instead of considering merely *a* as contrasted with (1) we might want a comparison of the yields of plots containing *a*, with or without any other ingredient, with the yields of plots not containing a at all, that is,

$$abc + ab + ac + a - bc - b - c - (1)$$

Or we might want a comparison of the effect of a in the presence of b with that of a in the absence of b, viz.,

$$abc + ab - ac - a - bc - b + c + (1)$$

This is also a comparison of the effect of b in the presence of a with that of b in the absence of a. We can thus arrive at the seven independent comparisons:

$$\begin{aligned}
A &= (a-1)(b+1)(c+1) = abc+ab+ac-bc+a-b-c-(1)\\
B &= (a+1)(b-1)(c+1) = abc+ab-ac+bc-a+b-c-(1)\\
C &= (a+1)(b+1)(c-1) = abc-ab+ac+bc-a-b+c-(1)\\
AB &= (a-1)(b-1)(c+1) = abc+ab-ac-bc-a-b+c+(1)\\
AC &= (a-1)(b+1)(c-1) = abc-ab+ac-bc-a+b-c+(1)\\
BC &= (a+1)(b-1)(c-1) = abc-ab-ac+bc+a-b-c+(1)\\
ABC &= (a-1)(b-1)(c-1) = abc-ab-ac-bc+a+b+c-(1)
\end{aligned}$$

Note the method of obtaining each of them as a symbolic algebraic product. Note also that they are orthogonal among themselves and likewise to the blocks,* so that the sum of their squares divided by the sum of the squares of their coefficients is equal to the sum of squares of their deviations from their mean. To each of them corresponds one degree of freedom, and the seven degrees of freedom due to treatments are completely accounted for in a manner which has a useful interpretation in treatment contrast.

As a numerical case, suppose that the total yields for five blocks are as follows, the unit being immaterial:

$$\begin{array}{cccc}
abc = 48 & ab = 54 & ac = 37 & bc = 20\\
a = 40 & b = 31 & c = 25 & (1) = 15
\end{array}$$

* Since the total yield of a block is

$$abc + ab + ac + bc + a + b + c + (1)$$

Then for the sum of squares of deviations for treatments we should have

$$\begin{aligned}&\tfrac{1}{5}[(48)^2 + (54)^2 + (37)^2 + (20)^2 + (40)^2 + (31)^2 + (25)^2 + (15)^2]\\ &\qquad - \frac{1}{5 \times 8}(48 + 54 + 37 + 20 + 40 + 31 + 25 + 15)^2\\ &\qquad = \tfrac{1}{5} \times 10400 - \tfrac{1}{40}(270)^2 = 2080 - 1822.5 = 257.5\end{aligned}$$

We take one-fifth of the square bracket because each total such as *abc* represents five blocks. Similarly, we take one-fortieth of the total squared since there are five blocks of eight plots each. Cf. formula (33) of section 63.

We also find

$$\begin{aligned}A &= 48 + 54 + 37 - 20 + 40 - 31 - 25 - 15 = 88\\ B &= 48 + 54 - 37 + 20 - 40 + 31 - 25 - 15 = 36\\ C &= 48 - 54 + 37 + 20 - 40 - 31 + 25 - 15 = -10\\ AB &= 48 + 54 - 37 - 20 - 40 - 31 + 25 + 15 = 14\\ AC &= 48 - 54 + 37 - 20 - 40 + 31 - 25 + 15 = -8\\ BC &= 48 - 54 - 37 + 20 + 40 - 31 - 25 + 15 = -24\\ ABC &= 48 - 54 - 37 - 20 + 40 + 31 + 25 - 15 = 18\end{aligned}$$

The sum of squares of these quantities is

$$(88)^2 + (36)^2 + (-10)^2 + (14)^2 + (-8)^2 + (-24)^2 + (18)^2 = 10{,}300$$

The sum of squares of coefficients of the *abc*, *ab*, etc., in any expression such as A is 8, and since each *abc*, *ab*, etc., is the total of 5 blocks we must divide 10,300 by 5 × 8. This gives 257.5, which agrees with the value obtained above.

69. Confounding. The treatment contrasts A, B, C are called *main effects*; the contrasts AB, AC, BC are called *interactions* of first order; ABC is an interaction of second order. If more ingredients are used or if different levels of ingredients are employed—double doses or triple doses—the number of separate treatments is greatly increased, and to accommodate all of them in each block would require blocks of large size, since the individual plots of the blocks can not be too small.* But if the blocks were larger we should be more likely to encounter soil heterogeneity within blocks. One way by which this can be controlled is by *confounding,* that is, by not completely replicating within each block. A simple example will make this clear.

Suppose that in the experiment just described we use ten blocks instead of five but have only four plots in each block. If five of the blocks contain the treatments

$$abc,\ a,\ b,\ c$$

and the other five the treatments

$$ab,\ ac,\ bc,\ (1)$$

the second-order interaction ABC is confounded with blocks. This interaction is not orthogonal to the blocks. For we recall that

$$ABC = abc - ab - ac - bc + a + b + c - (1)$$

and this is orthogonal neither to

$$abc + a + b + c$$

nor to

$$ab + ac + bc + (1)$$

The main effects and the first-order interactions are orthogonal to the blocks and are unaffected by block differences, since for any one of these there are two positive and two negative treatment combinations occurring in each block. The degrees of freedom for the analysis of variance would appear as follows:

	Degrees of freedom
Blocks.	$10 - 1 = 9$
Treatments (A, B, C, AB, AC, BC) . . .	6
Error.	$2(5 - 1)(4 - 1) = 24$
Total.	$4 \times 10 - 1 = 39$

* The language used is that of agricultural experiment, but it seems best to speak in concrete terms belonging to a specific science rather than to attempt to invent terms of greater generality, which, although of wider applicability, might not convey the concepts so well.

The contrast ABC has been sacrificed but it has been determined experimentally that higher-order interactions are often unimportant. In this case it is possible to evaluate the other six contrasts with whatever additional precision has been gained by eliminating soil heterogeneity through the use of smaller blocks.

70. Partial confounding. We have seen that, in an experiment involving the ingredients a, b, c in which a treatment may contain none, one, or two of these ingredients, it is possible to confound one of the treatment effects such as ABC with blocks, thus gaining greater precision by using homogeneous material, at the expense, however, of losing all information concerning the effect ABC. Instead of confounding this same interaction in all the blocks it would be possible to confound different interactions in different sets of blocks. This procedure is called *partial confounding*. Some information will then be obtained on all the interactions, the loss of information being spread over several interactions instead of being confined to one.

In the foregoing example we might confound each of the first-order interactions AB, AC, BC in a quarter of the blocks and the second-order interaction ABC in the remaining quarter. A complete set of blocks would then look like the accompanying diagram.

abc	ab	abc	bc	abc	ac	abc	a
c	(1)	a	(1)	b	(1)	b	c
ac	bc	ab	ac	ab	bc	ab	ac
a	b	b	c	a	c	bc	(1)

Here the interaction AB is confounded in the first column of blocks, BC in the second column, CA in the third, and ABC in the fourth.

To illustrate the use of the analysis of variance in partial confounding let us take an even simpler example, one in which we deal with only two ingredients a and b. We remember that the main effects and the interaction are

$$A = (a - 1)(b + 1) = ab + a - b - (1)$$

$$B = (a + 1)(b - 1) = ab + b - a - (1)$$

$$AB = (a - 1)(b - 1) = ab + (1) - a - b$$

Suppose that we confound A in half of the blocks and B in the other half, leaving AB unconfounded With eight blocks we could have two replications as shown below.*

I	II	III	IV
ab	b	ab	a
a	(1)	b	(1)

I′	II′	III′	IV′
ab	b	ab	a
a	(1)	b	(1)

We note that A is confounded in blocks I, II, I′, II′, B in blocks III, IV, III′, IV′, and that AB is unconfounded, since from each block we use a plus sign with one treatment and a minus sign with the other.

We can then calculate the AB sum of squares from all the blocks. A is calculated from blocks III, IV, III′, IV′; B, from blocks I, II, I′, II′. The sum of squares for error can be obtained by contrasting treatment differences in blocks I and I′, II and II′, III and III′, IV and IV′. These contrasts give us four degrees of freedom; a fifth degree of freedom for error is involved in the error component obtained when abstracting AB. The sum of squares for blocks (seven degrees of freedom) can be obtained directly or may be resolved into contrasts between I and I′, II and II′, III and III′, IV and IV′ (one degree in each of these, or four degrees altogether); the contrast between I + II + I′ + II′ and III + IV + III′ + IV′ (one degree) which is the blocks component obtained when evaluating AB; I + I′ versus II + II′

* They are not randomized in the diagram.

(one degree); and III + III′ versus IV + IV′ (one degree). These last two contrasts are seen to be closely related to the treatment effects A and B, being the block differences with which they are partially confounded.

The degrees of freedom table for the analysis of variance would then appear as follows:

	Degrees of Freedom
Blocks	7
Treatments (A, B, AB)	3
Error	5
Total	15

This section will be concluded by a simple numerical illustration of the set-up just described. Suppose that the values appertaining to the plots are as shown below:

I	II	III	IV
ab 7	*b* 2	*ab* 9	*a* 8
a 5	(1) 2	*b* 4	(1) 2
12	4	13	10

I′	II′	III′	IV′
ab 6	*b* 3	*ab* 7	*a* 6
a 4	(1) 2	*b* 3	(1) 2
10	5	10	8

The total sum of squares of deviations is

$$7^2+5^2+2^2+2^2+9^2+4^2+8^2+2^2+6^2+4^2+3^2+2^2+7^2+3^2+6^2+2^2$$

$$-\tfrac{1}{16}(7+5+2+2+9+4+8+2+6+4+3+2+7+3+6+2)^2$$

$$= 410 - \frac{(72)^2}{16} = 410 - 324 = 86$$

For blocks we find

$$\tfrac{1}{2}[(12)^2 + 4^2 + (13)^2 + (10)^2 + (10)^2 + 5^2 + (10)^2 + 8^2] - \frac{(72)^2}{16}$$

$$=359 - 324 = 35$$

As stated above, the AB sum of squares is calculated from all blocks. So as to be able to obtain the error component when abstracting AB, we make the following arrangement (Table 55):

TABLE 55

Treatments	Blocks		Total
	I, II I′, II′	III, IV III′, IV′	
$ab + (1)$	17	20	37
$a + b$	14	21	35
Total	31	41	72

The sum of squares of deviations can be calculated in the ordinary manner, but we shall here make use of the following formulas, which are applicable when we have only two divisions, and which are simpler from the standpoint of computation. Consider Table 56. The sums of squares of deviations are:

TABLE 56

		Total
	X_{11} X_{12}	T_1
	X_{21} X_{22}	T_2
Total	T'_1 T'_2	T

For rows $\frac{1}{4}(T_1 - T_2)^2$

For columns $\frac{1}{4}(T'_1 - T'_2)^2$

For error $\frac{1}{4}[(X_{11} + X_{22})^2 - (X_{12} + X_{21})^2]$

The first two of these follow from (1) of section 66; the second is readily demonstrated. In applying them to Table 55 we must realize that each difference of totals involves sixteen plots instead of four and that consequently we must multiply by $\frac{1}{16}$ instead of $\frac{1}{4}$.

TABLE 57

	Sum of squares of deviations	Degrees of freedom
Treatment AB	$\frac{1}{16}(37-35)^2=0\ 25$	1
Blocks $\left\{\begin{matrix} (\mathrm{I}+\mathrm{II}+\mathrm{I}'+\mathrm{II}') \\ -(\mathrm{III}+\mathrm{IV}+\mathrm{III}'+\mathrm{IV}') \end{matrix}\right\}$	$\frac{1}{16}(31-41)^2=6\ 25$	1
Error .	$\frac{1}{16}[(17+21)-(20+14)]^2=1$	1
Total	7 5	3

To check the total we calculate

$$\tfrac{1}{4}[(17)^2 + (20)^2 + (14)^2 + (21)^2] - \tfrac{1}{16}(72)^2 = 7.5$$

To calculate the sum of squares of deviations for treatment contrast A, we use only blocks III, IV, III′, IV′, since this contrast is confounded in the others. We find this sum to be equal to

$$\tfrac{1}{8}[ab + a - b - (1)]^2 = \tfrac{1}{8}(9 + 7 + 8 + 6 - 4 - 3 - 2 - 2)^2$$
$$= \tfrac{1}{8}(19)^2 = 45.125$$

From blocks I, II, I′, II′ we find the sum of squares of deviations for treatment contrast B to be equal to

$$\tfrac{1}{8}[ab - a + b - (1)]^2 = \tfrac{1}{8}(7 + 6 - 5 - 4 + 2 + 3 - 2 - 2)^2$$
$$= \tfrac{1}{8} \times 5^2 = 3.125$$

We are now ready to abstract the remaining sums of squares of deviations for error. If there is a fertility difference between two similar blocks, such as III and III′, it will be eliminated if we compare the differences in yield due to treatment, viz., $ab - b$, in these two blocks. These differences are $9 - 4 = 5$, and $7 - 3 = 4$, respectively. The sum of squares of deviations can be calculated either as

$$\tfrac{1}{2}(5^2 + 4^2) - \tfrac{1}{4}(5 + 4)^2 = 0.25$$

or as $(5 - 4)^2/4 = 0.25$, the latter way being in general simpler.

The treatment differences for the various blocks are shown below.

TABLE 58A

	$ab - a$
I	$7-5=2$
I′	$6-4=2$
Difference =0	

TABLE 58B

	$b - (1)$
II	$2-2=\ 0$
II′	$3-2=\ 1$
Difference = −1	

TABLE 58C

	$ab - b$
III	$9-4=5$
III′	$7-3=4$
Difference =1	

TABLE 58D

	$a - (1)$
IV	$8-2=6$
IV′	$6-2=4$
Difference =2	

TABLE 58E

	$ab - a - b + (1)$
I + II + I′+ II′	$13 - 9 - 5 + 4 = 3$
III + IV + III′ + IV′	$16 - 14 - 7 + 4 = -1$
	Difference = 4

Thus for the sum of squares of deviations due to error we have the following analysis:

TABLE 59

Error

Blocks	Sum of squares of deviations	Degrees of freedom
I − I′	$\frac{1}{4} \times 0^2 = 0$	1
II − II′	$\frac{1}{4} \times (-1)^2 = 0\ 25$	1
III − III′	$\frac{1}{4} \times 1^2 = 0\ 25$	1
IV − IV′	$\frac{1}{4} \times 2^2 = 1$	1
(I + II + I′ + II′) −(III + IV + III′ + IV′)	(See *AB* analysis) 1	1
Total	2 5	5

Although the sum of squares of deviations among block means has already been calculated we shall for completeness show the analysis.

TABLE 60

Blocks

	Sum of squares of deviations	Degrees of freedom
I − I′	$\frac{1}{4}(12 - 10)^2$ = 1	1
II − II′	$\frac{1}{4}(4 - 5)^2$ = 0 25	1
III − III′	$\frac{1}{4}(13 - 10)^2$ = 2 25	1
IV − IV′	$\frac{1}{4}(10 - 8)^2$ = 1	1
(I + II + I′ + II′) − (III + IV + III′ + IV′)	(See *AB* analysis) 1	1
(I + I′) − (II + II′)	$\frac{1}{8}(22 - 9)^2$ = 21 125	1
(III + III′) − (IV + IV′)	$\frac{1}{8}(23 - 18)^2$ = 3 125	1
Total	35	7

Results are recapitulated in Table 61.

TABLE 61

Analysis of Variance for Partial Confounding

		Sum of squares of deviations		Degrees of freedom	
Blocks			35		7
Treatments	*AB*	0 25	48 5	1	3
	A	45 125		1	
	B	3 125		1	
Error			2 5		5
Total			86		15

71. Dummy treatments. Sometimes, in order to have each ingredient or factor occur with proportional frequency in combination with the variants of other factors, it becomes necessary to use so-called *dummy treatments*. For example, if we are investigating three levels (degrees) and three kinds of nitrogenous fertilizer, the three plots of each replication receiving no nitrogen are indistinguishable. The accompanying diagram, representing such a replication, will make this clearer. The subscripts represent the level of the ingredient, that is, the quantities of it applied. The subscript 0 denotes that none has been applied, the subscript 1 indicates a single application, the subscript 2 indicates a double dose. The primes, etc., represent different kinds of ingredients; for example, they might be sulphate of ammonia, chloride of ammonia, and cyanamide. It is evident that there is no difference in the a_0's; in fact, their superscripts should be deleted since they are meaningless. Thus we have not nine distinct treatments but only seven, and our degrees of freedom in the analysis of variance would be, if we had, say, five blocks:

a_0'	a_0''	a_0'''
a_1'	a_1''	a_1'''
a_2'	a_2''	a_2'''

		Degrees of Freedom
Blocks		$5 - 1 = 4$
Treatments		$7 - 1 = 6$
Error	Among blocks	$4 \times 6 = 24$
	Within blocks	$5(3 - 1) = 10$
Total		$5 \times 9 - 1 = 44$

If we wish to analyze the treatment degrees of freedom further we may do so as follows: two degrees for comparison of the three levels of a, i.e., a_0, a_1, a_2; two degrees for comparison of the three kinds of treatment, i.e., a', a'', a'''; two degrees, $(3 - 1)(2 - 1)$, for interaction between level and kind, i.e.,

$$\begin{array}{ccc} a_1' & a_1'' & a_1''' \\ a_2' & a_2'' & a_2''' \end{array}$$

(The zero level would not be included here.)

The sum of squares of deviations for error within blocks is obtained by comparing the three dummy plots in each block. There are thus two degrees of freedom for each block, or ten altogether.

The sum of squares of deviations for error among blocks is composed of three parts:

(*i*) The remainder or interaction term obtained when comparing the three treatments a_1', a_1'', a_1''' in the five blocks, the degrees of freedom being $(3 - 1)(5 - 1) = 8$.

(*ii*) The interaction term obtained when comparing the three treatments a_2', a_2'', a_2''' in the five blocks, the degrees of freedom being eight as before.

(*iii*) The interaction term obtained when comparing the three levels of treatment a_0, a_1, a_2 in the five blocks, the number of degrees of freedom again being eight.

It is quite possible to have dummy treatments in connection with confounding, but for this and for more complex examples the reader is referred to Fisher * and Yates.†

As a numerical example of dummy treatments consider the following two blocks, where the notation has the same meaning as earlier in the section; that is, subscripts refer to different levels of treatment and superscripts to different kinds of treatment. The numbers are yields.

BLOCK I

			Total
a_0 2	a_0 4	a_0 2	8
a_1' 6	a_1'' 4	a_1''' 3	13
a_2' 7	a_2'' 5	a_2''' 9	21
			— 42

BLOCK II

			Total
a_0 3	a_0 2	a_0 2	7
a_1' 7	a_1'' 6	a_1''' 5	18
a_2' 8	a_2'' 6	a_2''' 9	23
			— 48

* R. A. Fisher, "The Design of Experiments"

† F. Yates, "The principles of orthogonality and confounding in replicated experiments," *Journal of Agricultural Science*, vol. 23, 1933, pp 108–144; "Complex experiments," *Supplement to the Journal of the Royal Statistical Society*, vol. 2, 1935, pp. 181–247.

The plots are not arranged at random, as they would be in the field, but are placed in corresponding positions so that totals of corresponding treatments can more readily be obtained.

We first calculate the total sum of squares of deviations, which is

$$2^2 + 4^2 + 2^2 + 6^2 + 4^2 + 3^2 + 7^2 + 5^2 + 9^2$$
$$+ 3^2 + 2^2 + 2^2 + 7^2 + 6^2 + 5^2 + 8^2 + 6^2 + 9^2 - (42 + 48)^2/18$$
$$= 548 - 450 = 98$$

The number of degrees of freedom is the number of plots less one, viz., 17.

For the difference between blocks we have

$$\frac{(42 - 48)^2}{18} = 2$$

with one degree of freedom.

We now take up the three different levels of treatment:

TABLE 62A

	a_0	a_1	a_2	Total
Block I ...	8	13	21	42
Block II	7	18	23	48
Total	15	31	44	90

The total sum of squares of deviations is

$$\frac{1}{3}\,[(8)^2 + (13)^2 + (21)^2 + (7)^2 + (18)^2 + (23)^2] - \frac{(90)^2}{18}$$
$$= 525\ \dot{3} - 4\dot{5}0 = 75.\dot{3}$$

with five degrees of freedom.

The sum of squares of deviations for different levels of treatment is

$$\frac{1}{6}\,[(15)^2 + (31)^2 + (44)^2] - \frac{(90)^2}{18} = 520.\dot{3} - 450 = 70.\dot{3}$$

with two degrees of freedom.

For block differences we have already found the sum of squares of deviations to be two, with one degree of freedom, and we can find the error term by subtraction:

$$75.\dot{3} - (70.\dot{3} + 2) = 3$$

the number of degrees of freedom being two.

Next we consider the different kinds of treatment. Here the a_0 plots are, of course, left out of consideration. Since we wish to get the interaction between levels and kinds of treatment we form the following table, in which blocks I and II are combined:

TABLE 62B

	a'	a''	a'''	Total
a_1	13	10	8	31
a_2	15	11	18	44
Total	28	21	26	75

The total sum of squares of deviations for this group is

$$\frac{1}{2}[(13)^2 + (10)^2 + (8)^2 + (15)^2 + (11)^2 + (18)^2] - \frac{(75)^2}{12}$$
$$= 501.5 - 468\ 75 = 32.75$$

the number of degrees of freedom being five. Note that each value, such as 13, is the total of two plot-yields, so that the total of 75 is for 12 plots.

For the sum of squares of deviations due to different kinds of treatment we obtain

$$\frac{1}{4}[(28)^2 + (21)^2 + (26)^2] - \frac{(75)^2}{12} = 475.25 - 468.75 = 6.5$$

with two degrees of freedom.

For the different levels in this group we find $(31 - 44)^2/12 = 14.083$, with one degree of freedom, so that for the interaction term among the three kinds and the two levels of treatment we find

$$32.75 - (6.5 + 14\ 08\dot{3}) = 12\ 1\dot{6}$$

The corresponding number of degrees of freedom is $(3 - 1)(2 - 1) = 2$, or it can be found by subtraction: $5 - (2 + 1) = 2$

This completes the analysis of the sum of squares due to treatments, but we found only part of the error term, viz , the interaction between blocks and levels of treatments (three, with two degrees of freedom). Consequently we next proceed to isolate the error or interaction term from blocks and kinds of treatment at the first level. (See Table 62C.)

TABLE 62C

	a_1'	a_1''	a_1'''	Total
Block I	6	4	3	13
Block II	7	6	5	18
Total	13	10	8	31

The sums of squares of deviations for this subgroup are:

Total: $6^2 + 4^2 + 3^2 + 7^2 + 6^2 + 5^2 - (31)^2/6$

$= 171 - 160.1\dot{6} = 10.8\dot{3}$ (5 degrees of freedom)

Treatment: $\frac{1}{2}[(13)^2 + (10)^2 + (8)^2] - (31)^2/6$

$= 166.5 - 160.1\dot{6} = 6.\dot{3}$ (2 degrees of freedom)

Blocks: $(13 - 18)^2/6 = 4.1\dot{6}$ (1 degree of freedom)

Error: $10.8\dot{3} - (6.\dot{3} + 4.1\dot{6})$

$= 0.\dot{3}$ (2 degrees of freedom)

Similarly we isolate the error term from blocks and kinds of treatment at the second level.

TABLE 62D

	a_2'	a_2''	a_2'''	Total
Block I .	7	5	9	21
Block II	8	6	9	23
Total .	15	11	18	44

Total: $7^2 + 5^2 + 9^2 + 8^2 + 6^2 + 9^2 - (44)^2/6$

$= 336 - 322.\dot{6} = 13.\dot{3}$ (5 degrees of freedom

Treatment: $\frac{1}{2}[(15)^2 + (11)^2 + (18)^2] - (44)^2/6$

$= 335 - 322.\dot{6} = 12.\dot{3}$ (2 degrees of freedom

Blocks: $(21 - 23)^2/6 = 0.6$ (1 degree of freedom)

Error: $13.3 - (12.3 + 0.6)$

$= 0.\dot{3}$ (2 degrees of freedom

The components of error between blocks may be summarize as follows:

TABLE 62E

	Sum of squares of deviations	Degrees of freedom
Blocks × levels of treatment	3	2
Blocks × kinds at 1st level.	0 3	2
Blocks × kinds at 2nd level	$0\ \dot{3}$	2
Between blocks . . .	$3\ \dot{6}$	6

It now remains only to find the components of error in the dummy treatments a_0 within blocks. These are

Block I $2^2 + 4^2 + 2^2 - 8^2/3 = 2\ \dot{6}$ (2 degrees of freedom)

Block II $3^2 + 2^2 + 2^2 - 7^2/3 = 0\ \dot{6}$ (2 degrees of freedom)

Total within blocks $3\ \dot{3}$ (4 degrees of freedom)

Collecting our results we have the following analysis:

TABLE 63
ANALYSIS OF VARIANCE FOR DUMMY TREATMENTS

	Sum of squares of deviations		Degrees of freedom	
Blocks		2		1
Treatments				
Levels .	70 3	89	2	6
Kinds	6 5		2	
Levels × kinds	12 1$\dot{6}$		2	
Error				
Between blocks. .	3 $\dot{6}$	7	6	10
Within blocks	3 $\dot{3}$		4	
Total		98		17

72. Non-orthogonal data. Non-orthogonality is sometimes deliberately introduced into an experimental design, as for example when confounding is resorted to. On the other hand, it may be unavoidable on account of the nature of the material, as for instance in poultry experiments in which the sex of the individual birds can not be determined when the experiment is begun, and when consequently the numbers of the different sexes in the various subclasses are not equal or not even proportional. When some of the animals in an experiment die during the progress of the experiment, or when some of the plots in a field trial are damaged, orthogonality is lost.

Some of the modifications of the ordinary analysis of variance for the case of deliberately non-orthogonal data have been described in the earlier sections of this chapter. No attempt will be made to discuss other types of non-orthogonal data, but attention is called to several original papers dealing with the subject.*

*S S Wilks, "The analysis of variance and covariance in non-orthogonal data," *Metron*, vol. 13, 1938, pp 141–154

F. Yates, "The principles of orthogonality and confounding in replicated experiments," *Journal of Agricultural Science*, vol 23, 1933, pp. 108–145; "The analysis of replicated experiments when the field results are incomplete," *Empire Journal of Experimental Agriculture*, vol 1, 1933, 129–142; "The analysis of multiple classifications with unequal numbers in the different classes," *Journal of the American Statistical Association*, vol. 29, 1934, pp. 51–66.

EXERCISES

1. Complete the analysis of variance for Table 52, p. 165.

2. Figure 14 shows the yields, in pounds per plot, in an experiment in raising potatoes Six different treatments were used in each of 4 different blocks The random field arrangement is shown in the table, in which t_i indicates the treatment applied to the particular plot. Analyze the variance, and make the appropriate tests.

Block I	t_2 306	t_4 442	t_5 295	t_1 290	t_6 457	t_3 349
Block II	t_1 253	t_6 415	t_3 297	t_2 288	t_5 268	t_4 434
Block III	t_4 419	t_2 307	t_1 178	t_5 310	t_6 467	t_3 304
Block IV	t_5 166	t_6 428	t_3 308	t_4 404	t_1 172	t_2 268

FIG. 14 —Yields of Potatoes in Pounds per Plot (Randomized Blocks)

3. Figure 15 shows the actual arrangement of a Latin square used in a potato-growing experiment in which 4 different treatments, t_1, t_2, t_3, t_4, were applied. The numbers in the various cells are yields in pounds per plot. (Eden and Fisher, *Journal of Agricultural Science,* vol. 19) Analyze the variance, and make a test of significance

t_3 444	t_4 422	t_1 173	t_2 398
t_1 279	t_2 439	t_3 423	t_4 409
t_4 436	t_3 428	t_2 445	t_1 212
t_2 453	t_1 237	t_4 410	t_3 393

FIG. 15 —Yields of Potatoes in Pounds per Plot (Latin Square)

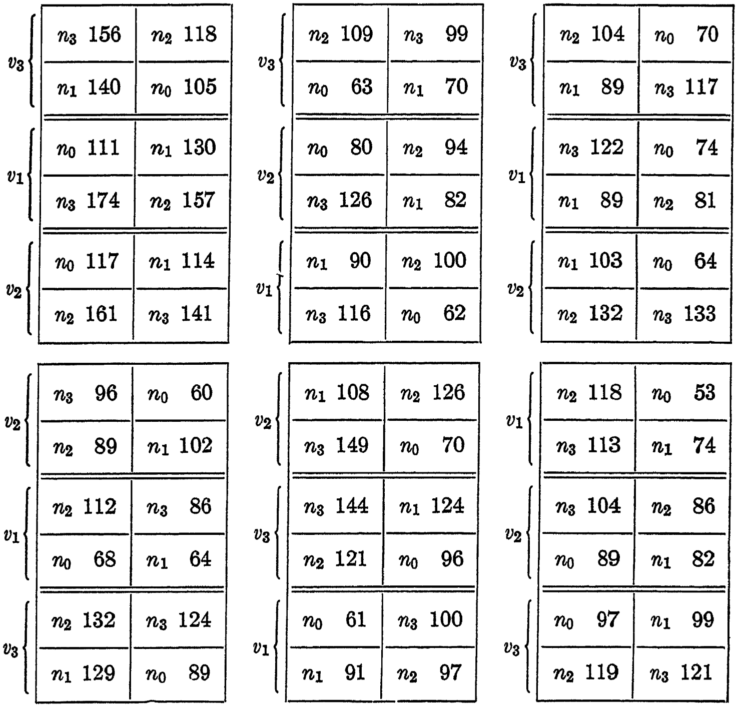
4. Figure 16 shows the plan, and the yield in quarter pounds, of an experiment in growing oats (Yates, *Supplement to the Journal of the Royal Statistical Society*, vol 2). There were 3 varieties, v_1, v_2, v_3, and 4 treatments, the latter consisting of the application of 4 different levels of nitrogenous fertilizer, n_0, n_1, n_2, n_3 Each block consisted of 3 whole plots, each of which contained 1 of the 3 varieties Each whole plot was subdivided into 4 subplots, each subplot containing a different level of nitrogen Such a design is called a *split-plot* arrangement Analyze the variance according to the following scheme, and make appropriate tests of significance:

Whole plots $\begin{cases}\text{Blocks} \\ \text{Varieties} \\ \text{Error}\end{cases}$

Subtotal

Subplots $\begin{cases}\text{Nitrogen} \\ \text{Nitrogen} \times \text{Varieties} \\ \text{Error}\end{cases}$

Total



v_3	n_3 156	n_2 118	v_3	n_2 109	n_3 99	v_3	n_2 104	n_0 70
	n_1 140	n_0 105		n_0 63	n_1 70		n_1 89	n_3 117
v_1	n_0 111	n_1 130	v_2	n_0 80	n_2 94	v_1	n_3 122	n_0 74
	n_3 174	n_2 157		n_3 126	n_1 82		n_1 89	n_2 81
v_2	n_0 117	n_1 114	v_1	n_1 90	n_2 100	v_2	n_1 103	n_0 64
	n_2 161	n_3 141		n_3 116	n_0 62		n_2 132	n_3 133

v_2	n_3 96	n_0 60	v_2	n_1 108	n_2 126	v_1	n_2 118	n_0 53
	n_2 89	n_1 102		n_3 149	n_0 70		n_3 113	n_1 74
v_1	n_2 112	n_3 86	v_3	n_3 144	n_1 124	v_2	n_3 104	n_2 86
	n_0 68	n_1 64		n_2 121	n_0 96		n_0 89	n_1 82
v_3	n_2 132	n_3 124	v_1	n_0 61	n_3 100	v_3	n_0 97	n_1 99
	n_1 129	n_0 89		n_1 91	n_2 97		n_2 119	n_3 121

FIG 16 —Experiment on Oats. Plan, and Yields in Quarter-Pounds

5. In Fig 17 is shown the design of an experiment on peas (Yates, *Supplement to the Journal of the Royal Statistical Society*, vol. 2). Three different kinds of fertilizer were used. nitrogen (*n*), phosphate (*p*), and potash (*k*) These were administered singly and in various combinations, including no fertilizer, which is indicated by the symbol (1) Half of the 6 blocks contained the treatments (1), *np*, *nk*, *pk*; the other half, the treatments *n*, *p*, *k*, *npk*. The yields of the various plots are given in pounds Perform an analysis of variance, and make any tests of significance that seem appropriate

npk 55 8	*p* 62 8	*npk* 58 5	*n* 59 8	*np* 62 8	*pk* 49 5
k 55 0	*n* 69 5	*p* 56 0	*k* 55 5	*nk* 57 0	(1) 46 8
np 59 0	*nk* 57 2	(1) 51 5	*np* 52 0	*npk* 48 8	*n* 62 0
(1) 56 0	*pk* 53 2	*pk* 48 8	*nk* 49 8	*p* 44 2	*k* 45 5

FIG. 17.—Experiment on Peas. Plan, and Yields in Pounds

TABLES

TABLE I

Probabilities and Ordinates of the Normal Curve Corresponding to Given Deviations

x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$	x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$
0 00	5000	3989	0 50	3085	3521
0 01	4960	3989	0 51	3050	.3503
0 02	4920	3989	0 52	3015	.3485
0 03	4880	3988	0 53	2981	3467
0 04	4840	3986	0 54	2946	3448
0 05	.4801	3984	0 55	2912	3429
0 06	.4761	3982	0 56	2877	.3410
0 07	4721	3980	0 57	2843	3391
0 08	4681	3977	0 58	2810	3372
0 09	4641	3973	0 59	2776	3352
0 10	4602	3970	0 60	2743	3332
0 11	4562	3965	0 61	2709	3312
0 12	4522	3961	0 62	2676	3292
0 13	4483	3956	0 63	2643	3271
0 14	.4443	3951	0 64	2611	3251
0 15	4404	3945	0 65	2578	3230
0 16	4364	3939	0 66	2546	3209
0 17	4325	3932	0 67	2514	3187
0 18	4286	3925	0 68	2483	3166
0 19	4247	3918	0 69	2451	3144
0 20	4207	3910	0 70	2420	3123
0 21	4168	3902	0 71	.2389	3101
0 22	4129	3894	0 72	2358	3079
0 23	4090	3885	0 73	2327	3056
0 24	4052	3876	0 74	2296	3034
0 25	4013	3867	0 75	2266	3011
0 26	3974	3857	0 76	2236	2989
0 27	3936	3847	0 77	2206	2966
0 28	3897	3836	0 78	2177	2943
0 29	3859	3825	0 79	2148	2920
0 30	3821	3814	0 80	2119	2897
0 31	3783	3802	0 81	2090	2874
0 32	3745	3790	0 82	2061	2850
0 33	3707	3778	0 83	2033	2827
0 34	3669	3765	0 84	2005	2803
0 35	3632	3752	0 85	1977	2780
0 36	3594	3739	0 86	1949	2756
0 37	3557	3725	0 87	1922	2732
0 38	3520	3712	0 88	1894	2709
0 39	3483	3697	0 89	1867	2685
0 40	3446	3683	0 90	1841	2661
0 41	3409	3668	0 91	· 1814	2637
0 42	3372	3653	0 92	.1788	2613
0 43	3336	3637	0 93	1762	2589
0 44	3300	3621	0 94	1736	2565
0 45	3264	3605	0 95	1711	2541
0 46	.3228	3589	0 96	1685	2516
0 47	3192	3572	0 97	1660	2492
0 48	3156	3555	0 98	1635	2468
0 49	3121	3538	0 99	1611	2444

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE I—*Continued*

Probabilities and Ordinates of the Normal Curve Corresponding to Given Deviations

x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$	x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$
1 00	1587	2420	1 50	0668	1295
1 01	1562	2396	1 51	0655	1276
1 02	1539	2371	1 52	0643	1257
1 03	1515	2347	1 53	0630	1238
1 04	1492	2323	1 54	0618	1219
1 05	1469	2299	1 55	0606	1200
1 06	1446	2275	1 56	0594	1182
1 07	1423	2251	1 57	0582	1163
1 08	1401	2227	1 58	0571	1145
1 09	1379	2203	1 59	0559	1127
1 10	1357	2179	1 60	0548	1109
1 11	1335	2155	1 61	0537	1092
1 12	1314	2131	1 62	0526	1074
1 13	1292	2107	1 63	0516	1057
1 14	1271	2083	1 64	0505	1040
1 15	1251	2059	1 65	0495	1023
1 16	1230	2036	1 66	0485	1006
1 17	1210	2012	1 67	0475	0989
1 18	1190	1989	1 68	0465	0973
1.19	1170	1965	1 69	0455	0957
1 20	1151	1942	1 70	0446	0940
1 21	1131	1919	1 71	0436	0925
1 22	1112	1895	1 72	0427	0909
1.23	1093	1872	1 73	0418	0893
1 24	1075	1849	1 74	0409	0878
1 25	1056	1826	1 75	0401	0863
1 26	1038	1804	1 76	0392	0848
1 27	1020	1781	1 77	0384	0833
1 28	1003	1758	1 78	0375	0818
1.29	0985	1736	1 79	0367	0804
1 30	0968	1714	1 80	0359	0790
1 31	0951	1691	1 81	0351	0775
1 32	0934	1669	1 82	0344	0761
1 33	0918	1647	1 83	0336	0748
1 34	.0901	1626	1 84	0329	0734
1 35	0885	1604	1 85	0322	0721
1 36	0869	1582	1 86	0314	0707
1 37	0853	1561	1 87	0307	0694
1 38	0838	1539	1 88	0301	0681
1.39	0823	1518	1 89	0294	0669
1 40	0808	1497	1 90	0287	0656
1 41	0793	1476	1 91	0281	0644
1 42	0778	1456	1 92	0274	0632
1 43	0764	1435	1 93	0268	0620
1 44	.0749	1415	1 94	.0262	0608
1 45	0735	1394	1 95	0256	0596
1 46	0721	1374	1 96	0250	0584
1 47	0708	1354	1 97	0244	0573
1 48	0694	1334	1 98	0239	0562
1 49	0681	1315	1 99	0233	0551

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE I—*Continued*

Probabilities and Ordinates of the Normal Curve Corresponding to Given Deviations

x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$	x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$
2 00	0228	0540	2 50	0062	0175
2 01	0222	0529	2 51	0060	0171
2 02	0217	0519	2 52	0059	0167
2 03	0212	0508	2 53	0057	0163
2 04	0207	0498	2 54	0055	0158
2 05	0202	.0488	2 55	0054	0154
2 06	0197	0478	2 56	0052	0151
2 07	0192	0468	2 57	0051	0147
2 08	0188	0459	2 58	0049	0143
2 09	0183	0449	2 59	0048	0139
2 10	0179	0440	2 60	0047	0136
2 11	0174	0431	2 61	0045	0132
2 12	0170	0422	2 62	0041	0129
2 13	0166	0413	2 63	0043	0126
2 14	0162	0404	2 64	0041	0122
2 15	0158	0395	2 65	0040	0119
2 16	0154	0387	2 66	0039	0116
2 17	0150	0379	2 67	0038	0113
2 18	0146	0371	2 68	0037	0110
2 19	0143	0363	2 69	0036	0107
2 20	0139	0355	2 70	0035	0104
2 21	0136	0347	2 71	0034	0101
2 22	0132	0339	2 72	0033	0099
2 23	0129	0332	2 73	0032	0096
2 24	0125	0325	2 74	0031	0093
2 25	0122	0317	2 75	0030	0091
2 26	0119	0310	2 76	0029	0088
2 27	0116	0303	2 77	0028	0086
2 28	0113	0297	2 78	0027	0084
2 29	0110	0290	2 79	0026	0081
2 30	0107	0283	2 80	0026	0079
2 31	0104	0277	2 81	0025	0077
2 32	0102	0270	2 82	0024	0075
2 33	0099	0264	2 83	0023	0073
2 34	0096	0258	2 84	0023	0071
2 35	0094	0252	2 85	0022	0069
2 36	0091	0246	2 86	0021	0067
2 37	0089	0241	2 87	0021	0065
2 38	0087	0235	2 88	0020	0063
2 39	0084	0229	2 89	0019	0061
2 40	0082	0224	2 90	0019	0060
2 41	0080	0219	2 91	0018	0058
2 42	0078	0213	2 92	0018	0056
2 43	0075	0208	2 93	0017	0055
2 44	0073	0203	2 94	0016	0053
2 45	0071	0198	2 95	0016	0051
2 46	0069	0194	2 96	0015	0050
2 47	0068	0189	2 97	0015	0048
2 48	0066	0184	2 98	0014	0047
2 49	0064	0180	2 99	0014	0046

The probability of a deviation *numerically* greater than x is twice the probability given in the table

TABLE I—*Continued*

Probabilities and Ordinates of the Normal Curve Corresponding to Given Deviations

x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$	x	Probability of a deviation greater than x	Ordinate $\frac{e^{-x^2/2}}{(2\pi)^{1/2}}$
3 00	0013	0044	3 50	0002	0009
3 01	0013	0043	3 51	0002	0008
3 02	0013	0042	3 52	0002	0008
3 03	0012	0040	3 53	0002	0008
3 04	0012	0039	3 54	0002	0008
3 05	0011	0038	3 55	0002	0007
3 06	0011	0037	3 56	0002	0007
3 07	0011	0036	3 57	0002	0007
3 08	0010	0035	3 58	0002	0007
3 09	0010	0034	3.59	0002	0006
3 10	0010	0033	3 60	0002	0006
3 11	0009	0032	3 61	0002	0006
3 12	0009	0031	3 62	0001	0006
3 13	0009	0030	3 63	0001	0005
3 14	0008	0029	3 64	0001	0005
3 15	0008	0028	3 65	0001	0005
3 16	0008	0027	3 66	0001	0005
3 17	0008	0026	3 67	0001	0005
3 18	0007	0025	3 68	0001	0005
3 19	0007	0025	3 69	0001	0004
3 20	0007	0024	3 70	0001	0004
3 21	0007	0023	3 71	0001	0004
3 22	0006	0022	3 72	0001	0004
3 23	0006	0022	3 73	0001	0004
3 24	0006	0021	3 74	0001	0004
3 25	0006	0020	3 75	0001	0004
3 26	0006	0020	3 76	0001	0003
3 27	0005	0019	3 77	0001	0003
3 28	0005	0018	3 78	0001	0003
3 29	0005	0018	3 79	0001	0003
3 30	0005	0017	3 80	0001	0003
3 31	0005	0017	3 81	0001	0003
3 32	0005	0016	3 82	0001	0003
3 33	0004	0016	3 83	0001	0003
3 34	0004	0015	3 84	0001	0003
3 35	0004	0015	3 85	0001	0002
3 36	0004	0014	3 86	0001	0002
3 37	0004	0014	3 87	0001	0002
3.38	0004	0013	3 88	0001	0002
3 39	0003	0013	3 89	0001	0002
3 40	0003	0012	3 90	0000	0002
3 41	0003	0012	3 91	0000	0002
3 42	0003	0012	3 92	0000	0002
3 43	0003	0011	3 93	0000	0002
3 44	0003	0011	3 94	0000	0002
3 45	0003	0010	3 95	0000	0002
3 46	0003	0010	3 96	0000	0002
3 47	0003	0010	3 97	0000	0002
3 48	0003	0009	3 98	0000	0001
3 49	0002	0009	3 99	0000	0001

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE II

Deviations of the Normal Curve Corresponding to Given Probabilities

Probability of a deviation greater than x	x	Probability of a deviation greater than x	x	Probability of a deviation greater than x	x
.000	∞	175	9346	350	3853
.005	2 5758	180	9154	355	3719
.010	2 3263	185	8965	360	3585
.015	2 1701	190	8779	365	3451
.020	2 0537	.195	8596	.370	3319
.025	1 9600	200	8416	375	3186
.030	1 8808	205	8239	380	3055
.035	1 8119	210	8064	385	2924
.040	1 7507	215	7892	390	2793
.045	1 6954	220	7722	395	2663
.050	1 6449	225	7554	400	2533
.055	1 5982	230	7388	405	2404
.060	1 5548	235	7225	410	2275
.065	1 5141	240	7063	415	2147
.070	1 4758	245	6903	420	2019
.075	1 4395	250	6745	425	-1891
080	1 4051	255	6588	430	1764
.085	1 3722	260	6433	435	1637
090	1 3408	265	6280	440	1510
.095	1 3106	270	6128	445	1383
100	1 2816	275	5978	450	1257
.105	1 2536	280	5828	455	1130
.110	1 2265	285	5681	460	1004
.115	1 2004	290	5534	465	0878
.120	1 1750	295	5388	.470	0753
.125	1 1503	300	5244	.475	0627
.130	1 1264	305	5101	480	0502
.135	1 1031	310	4959	485	0376
140	1 0803	315	4817	.490	0251
.145	1 0581	320	4677	495	0125
150	1 0364	325	4538	500	0000
155	1 0152	330	4399		
160	9945	335	4261		
165	.9741	.340	.4125		
.170	.9542	.345	3989		

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE III

Probabilities of the Normal Curve Corresponding to Large Deviations

x	Probability of a deviation greater than x	x	Probability of a deviation greater than x
3	00134 99	6	$9\ 8660 \times 10^{-10}$
3 5	00023 26	7	$1\ 2798 \times 10^{-12}$
4	00003 17	8	$6\ 2210 \times 10^{-16}$
4 5	00000 33977	9	$1\ 1286 \times 10^{-19}$
5	00000 02867	10	$7\ 6199 \times 10^{-24}$

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE IV

Deviations of the Normal Curve Corresponding to Small Probabilities

Probability of a deviation greater than x	x	Probability of a deviation greater than x	x
005	2 57583	000,000,5	4 89164
000,5	3 29053	000,000,05	5 32672
000,05	3 89059	000,000,005	5 73073
000,005	4 41717	000,000,000,5	6 10941

The probability of a deviation *numerically* greater than x is twice the probability given in the table.

TABLE V

Values of t Corresponding to Given Probabilities *

Degrees of freedom n	Probability of a deviation greater than t					
	005	01	025	.05	.1	15
1	63 657	31 821	12 706	6 314	3 078	1 963
2	9 925	6 965	4 303	2 920	1 886	1 386
3	5 841	4 541	3 182	2 353	1 638	1 250
4	4 604	3 747	2 776	2 132	1 533	1 190
5	4 032	3 365	2 571	2 015	1 476	1 156
6	3 707	3 143	2 447	1 943	1 440	1 134
7	3 499	2 998	2 365	1 895	1 415	1 119
8	3 355	2 896	2 306	1 860	1 397	1 108
9	3 250	2 821	2 262	1 833	1 383	1 100
10	3.169	2 764	2 228	1 812	1 372	1 093
11	3 106	2 718	2 201	1 796	1 363	1 088
12	3 055	2 681	2 179	1 782	1 356	1 083
13	3 012	2 650	2 160	1 771	1 350	1 079
14	2 977	2 624	2 145	1 761	1 345	1 076
15	2 947	2 602	2 131	1 753	1 341	1 074
16	2 921	2 583	2 120	1 746	1 337	1 071
17	2 898	2 567	2 110	1 740	1 333	1 069
18	2 878	2 552	2 101	1 734	1 330	1 067
19	2 861	2 539	2 093	1 729	1 328	1 066
20	2 845	2 528	2 086	1 725	1 325	1 064
21	2 831	2 518	2 080	1 721	1 323	1 063
22	2 819	2 508	2 074	1 717	1 321	1 061
23	2 807	2 500	2 069	1 714	1 319	1 060
24	2 797	2 492	2 064	1 711	1 318	1 059
25	2 787	2 485	2 060	1 708	1 316	1 058
26	2 779	2 479	2 056	1 706	1 315	1 058
27	2 771	2 473	2 052	1 703	1 314	1 057
28	2 763	2.467	2 048	1 701	1 313	1 056
29	2 756	2 462	2 045	1 699	1 311	1 055
30	2 750	2.457	2 042	1 697	1 310	1 055
∞	2 576	2 326	1 960	1 645	1 282	1 036

The probability of a deviation *numerically* greater than t is twice the probability given at the head of the table.

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R A Fisher, and the publishers, Messrs. Oliver and Boyd

TABLE V—*Continued*

Values of t Corresponding to Given Probabilities

Degrees of freedom n	Probabılıty of a devıatıon greater than t					
	2	25	3	35	4	45
1	1 376	1 000	727	510	325	158
2	1 061	816	617	445	289	142
3	978	765	584	424	277	.137
4	941	741	569	414	271	.134
5	920	.727	559	408	267	132
6	906	.718	553	404	265	.131
7	896	.711	549	402	263	130
8	889	.706	546	399	262	130
9	883	.703	543	398	261	129
10	.879	.700	.542	.397	260	.129
11	876	.697	540	396	260	.129
12	873	.695	539	.395	259	.128
13	870	.694	538	394	259	128
14	868	.692	537	393	258	128
15	866	.691	.536	.393	258	128
16	865	690	535	392	258	.128
17	863	689	534	392	257	128
18	862	688	534	.392	257	127
19	861	.688	533	391	257	127
20	860	.687	533	391	257	.127
21	859	.686	532	.391	257	127
22	858	686	532	.390	256	.127
23	858	.685	532	390	256	.127
24	857	.685	531	390	256	.127
25	856	.684	531	390	.256	127
26	856	.684	.531	390	256	127
27	855	684	531	389	256	127
28	855	.683	530	389	256	.127
29	854	.683	.530	389	256	127
30	854	683	530	389	256	.127
∞	842	674	524	385	253	.126

The probabılity of a devıatıon *numerıcally* greater than t is twice the probabılity gıven at the head of the table.

TABLE VI

VALUES OF χ^2 CORRESPONDING TO GIVEN PROBABILITIES *

Degrees of freedom n	Probability of a deviation greater than χ^2						
	01	02	05	10	20	30	50
1	6 635	5 412	3 841	2 706	1 642	1 074	455
2	9 210	7 824	5 991	4 605	3 219	2 408	1 386
3	11 341	9 837	7 815	6 251	4 642	3 665	2 366
4	13 277	11 668	9 488	7 779	5 989	4 878	3 357
5	15 086	13 388	11 070	9 236	7 289	6 064	4 351
6	16 812	15 033	12 592	10 645	8 558	7 231	5 348
7	18 475	16 622	14 067	12 017	9 803	8 383	6 346
8	20 090	18 168	15 507	13 362	11 030	9 524	7 344
9	21 666	19 679	16 919	14 684	12 242	10 656	8 343
10	23 209	21 161	18.307	15 987	13 442	11 781	9 342
11	24 725	22 618	19 675	17 275	14 631	12 899	10 341
12	26 217	24 054	21 026	18.549	15 812	14 011	11 340
13	27 688	25 472	22 362	19 812	16 985	15 119	12 340
14	29 141	26 873	23 685	21 064	18 151	16 222	13 339
15	30 578	28 259	24 996	22 307	19 311	17 322	14 339
16	32 000	29 633	26 296	23 542	20 465	18 418	15 338
17	33 409	30 995	27 587	24 769	21 615	19 511	16 338
18	34 805	32 346	28 869	25 989	22 760	20 601	17 338
19	36 191	33 687	30 144	27 204	23 900	21 689	18 338
20	37 566	35.020	31 410	28 412	25 038	22 775	19 337
21	38 932	36 343	32 671	29 615	26 171	23 858	20 337
22	40 289	37 659	33 924	30 813	27 301	24 939	21 337
23	41 638	38 968	35 172	32 007	28 429	26 018	22 337
24	42 980	40 270	36.415	33 196	29 553	27 096	23 337
25	44 314	41 566	37 652	34 382	30 675	28 172	24 337
26	45 642	42 856	38 885	35 563	31.795	29.246	25 336
27	46 963	44 140	40 113	36 741	32 912	30 319	26 336
28	48 278	45 419	41 337	37 916	34 027	31 391	27 336
29	49 588	46 693	42.557	39.087	35 139	32 461	28 336
30	50 892	47 962	43 773	40 256	36 250	33 530	29 336

For larger values of n, the quantity $(2\chi^2)^{½} - (2n - 1)^{½}$ may be used as a normal deviate with unit standard deviation.

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R A Fisher, and the publishers, Messrs. Oliver and Boyd.

TABLE VI—*Continued*

VALUES OF χ^2 CORRESPONDING TO GIVEN PROBABILITIES

Degrees of freedom n	Probability of a deviation greater than χ^2					
	.70	80	90	95	98	99
1	.148	0642	0158	00393	000628	000157
2	.713	446	211	103	0404	0201
3	1.424	1 005	584	352	185	115
4	2 195	1 649	1 064	711	.429	297
5	3 000	2 343	1 610	1 145	752	554
6	3 828	3 070	2 204	1 635	1 134	872
7	4 671	3 822	2 833	2 167	1 564	1 239
8	5 527	4 594	3 490	2 733	2 032	1 646
9	6.393	5 380	4 168	3 325	2 532	2 088
10	7 267	6 179	4 865	3 940	3 059	2 558
11	8 148	6 989	5 578	4 575	3 609	3 053
12	9 034	7 807	6 304	5 226	4 178	3 571
13	9 926	8 634	7 042	5 892	4 765	4 107
14	10 821	9 467	7 790	6 571	5 368	4 660
15	11 721	10 307	8.547	7 261	5 985	5.229
16	12 624	11 152	9 312	7 962	6 614	5 812
17	13 531	12 002	10 085	8 672	7 255	6 408
18	14 440	12 857	10 865	9 390	7 906	7 015
19	15 352	13 716	11 651	10 117	8 567	7 633
20	16 266	14 578	12 443	10 851	9 237	8 260
21	17 182	15 445	13 240	11 591	9 915	8 897
22	18 101	16 314	14 041	12 338	10 600	9 542
23	19 021	17 187	14 848	13 091	11 293	10 196
24	19 943	18 062	15 659	13 848	11 992	10 856
25	20 867	18 940	16 473	14 611	12 697	11 524
26	21 792	19 820	17 292	15 379	13 409	12.198
27	22 719	20 703	18 114	16 151	14 125	12 879
28	23 647	21 588	18 939	16 928	14 847	13 565
29	24 577	22 475	19 768	17 708	15 574	14 256
30	25 508	23 364	20 599	18.493	16 306	14 953

For larger values of n, the quantity $(2\chi^2)^{1/2} - (2n - 1)^{1/2}$ may be used as a normal deviate with unit standard deviation.

TABLE VII

5 Per Cent Points of the Distribution of z *

Degrees of freedom n_2 of smaller mean square	Degrees of freedom n_1 of greater mean square					
	1	2	3	4	5	6
1	2 5421	2 6479	2 6870	2 7071	2 7194	2.7276
2	1 4592	1 4722	1.4765	1.4787	1 4800	1 4808
3	1 1577	1.1284	1.1137	1.1051	1 0994	1 0953
4	1 0212	9690	.9429	.9272	9168	9093
5	.9441	.8777	.8441	8236	8097	7997
6	8948	8188	7798	7558	7394	7274
7	8606	.7777	.7347	7080	.6896	6761
8	.8355	.7475	.7014	6725	6525	6378
9	8163	7242	6757	6450	6238	6080
10	.8012	.7058	6553	6232	6009	5843
11	·7889	.6909	6387	6055	5822	5648
12	.7788	6786	6250	5907	5666	5487
13	.7703	6682	6134	5783	5535	5350
14	.7630	6594	6036	5677	5423	5233
15	.7568	6518	.5950	5585	5326	5131
16	7514	6451	5876	5505	5241	5042
17	7466	.6393	5811	5434	5166	4964
18	7424	.6341	5753	5371	5099	4894
19	7386	.6295	5701	5315	5040	4832
20	7352	.6254	5654	5265	4986	4776
21	7322	6216	5612	5219	4938	4725
22	7294	6182	5574	5178	4894	4679
23	7269	6151	5540	5140	4854	4636
24	7246	6123	5508	5106	4817	4598
25	7225	6097	5478	.5074	4783	4562
26	.7205	6073	.5451	5045	4752	4529
27	.7187	6051	5427	5017	4723	4499
28	.7171	.6030	5403	4992	4696	4471
29	.7155	6011	5382	4969	4671	.4444
30	.7141	.5994	5362	4947	4648	4420
40	.7037	5866	5217	4789	.4479	4242
60	6933	5738	5073	.4632	4311	.4064
120	6830	5611	.4930	4475	4143	3885
∞	6729	5486	4787	4319	.3974	3706

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R A Fisher, and the publishers, Messrs Oliver and Boyd.

TABLE VII—*Continued*

5 Per Cent Points of the Distribution of z

Degrees of freedom n_2 of smaller mean square	Degrees of freedom n_1 of greater mean square					
	7	8	10	12	24	∞
1	2 7335	2 7380	2 7442	2 7484	2 7588	2 7693
2	1 4814	1.4819	1 4826	1 4830	1 4840	1 4851
3	1.0922	1 0899	1 0865	1 0842	1 0781	1 0716
4	9037	8993	8929	8885	8767	8639
5	7921	.7862	7775	7714	7550	7368
6	7184	7112	7006	6931	6729	6499
7	6658	6576	6455	6369	6134	.5862
8	6265	6175	6041	5945	5682	.5371
9	5959	5862	5717	5613	5324	4979
10	5714	5611	5457	.5346	5035	.4657
11	5514	5406	5243	.5126	.4795	.4387
12	5347	.5234	5064	.4941	4592	4156
13	5206	5089	.4912	.4785	.4419	3957
14	5084	.4964	4782	4649	.4269	3782
15	4979	4855	4668	4532	.4138	3628
16	4887	.4760	4568	4428	.4022	3490
17	4805	4676	4480	4337	3919	3366
18	4733	4602	4402	4255	3827	3253
19	4668	.4535	.4331	4182	3743	3151
20	4610	.4474	4268	4116	3668	.3057
21	4557	4420	4211	.4055	3599	2971
22	4509	4370	4158	4001	.3536	2892
23	4465	.4325	.4110	3950	3478	2818
24	.4425	4283	4066	.3904	.3425	2749
25	4388	4244	.4025	.3862	.3376	2685
26	.4354	4209	.3987	.3823	.3330	2625
27	4322	4176	.3952	3786	.3287	2569
28	.4292	4146	.3920	3752	.3248	2516
29	4265	4117	.3889	3720	3211	2466
30	.4239	4090	.3861	.3691	3176	2419
40	.4053	3897	3655	.3475	.2920	2047
60	3866	3702	3447	.3255	2654	1644
120	3678	.3506	.3236	3032	.2376	1131
∞	3490	.3309	3023	2804	2085	0000

TABLE VIII

1 Per Cent Points of the Distribution of z *

Degrees of freedom n_2 of smaller mean square	Degrees of freedom n_1 of greater mean square					
	1	2	3	4	5	6
1	4 1535	4 2585	4 2974	4 3175	4 3297	4 3379
2	2 2950	2 2976	2 2984	2 2988	2 2991	2 2992
3	1 7649	1 7140	1 6915	1 6786	1 6703	1 6645
4	1 5270	1 4452	1 4075	1 3856	1 3711	1 3609
5	1 3943	1.2929	1 .2449	1 2164	1 1974	1 1838
6	1 3103	1 1955	1 1401	1 1068	1 0843	1 0680
7	1 2526	1 1281	1 0672	1 0300	1 0048	9864
8	1 2106	1 0787	1 0135	9734	9459	.9259
9	1 1786	1 0411	9724	9299	9006	8791
10	1 1535	1 0114	9399	8954	8646	8419
11	1.1333	9874	9136	8674	8354	8116
12	1 1166	9677	.8919	8443	8111	7864
13	1 1027	9511	8737	8248	7907	7652
14	1 0909	.9370	.8581	.8082	7732	7471
15	1 0807	9249	8448	.7939	7582	7314
16	1 0719	9144	8331	7814	7450	7177
17	1 0641	9051	8229	7705	7335	7057
18	1 0572	8970	8138	7607	7232	6950
19	1 0511	.8897	8057	.7521	7140	6854
20	1 0457	8831	.7985	.7443	7058	.6768
21	1 0408	8772	.7920	7372	6984	6690
22	1 0363	.8719	.7860	7309	6916	6620
23	1 0322	8670	7806	.7251	6855	6555
24	1 0285	.8626	7757	7197	6799	6496
25	1.0251	.8585	7712	.7148	6747	.6442
26	1 0220	.8548	7670	7103	6699	6392
27	1 0191	.8513	7631	.7062	6655	6346
28	1 0164	.8481	7595	.7023	6614	6303
29	1 0139	.8451	7562	.6987	.6576	6263
30	1 0116	.8423	7531	.6954	6540	6226
40	9949	.8223	7307	.6712	6283	5956
60	9784	.8025	7086	6472	6028	5687
120	.9622	.7829	.6867	.6234	5774	5419
∞	.9462	.7636	6651	5999	.5522	5152

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R A Fisher, and the publishers, Messrs. Oliver and Boyd

TABLE VIII—*Continued*

1 Per Cent Points of the Distribution of z

Degrees of freedom n_2 of smaller mean square	Degrees of freedom n_1 of greater mean square					
	7	8	10	12	24	∞
1	4 3438	4 3482	4 3544	4 3585	4 3689	4 3794
2	2 2993	2 2994	2 2996	2 2997	2 2999	2 3001
3	1 6602	1 6569	1 6522	1 6489	1 6404	1 6314
4	1 3532	1 3473	1 3387	1 3327	1 3170	1 3000
5	1 1736	1 1656	1 1539	1 1457	1 1239	1 0997
6	1 0557	1 0460	1 0318	1 0218	9948	9643
7	9724	9614	9451	9335	9020	8658
8	.9105	.8983	.8802	8673	8319	7904
9	8626	8494	8297	8157	7769	7305
10	.8244	8104	7894	7744	7324	6816
11	.7932	.7785	7564	7405	6958	6408
12	.7673	7520	.7289	7122	6649	6061
13	.7454	7295	7056	6882	6386	5761
14	7267	7103	6856	.6675	6159	5500
15	.7105	6937	6682	6496	.5961	5269
16	6963	6791	6530	6339	5786	5064
17	6839	6663	6395	6199	5630	4879
18	6728	6549	6276	6075	5491	4712
19	6629	6447	6169	5964	5366	4560
20	6540	6355	6072	5864	5253	4421
21	6459	.6272	5984	5773	5150	4294
22	6386	6196	5904	5691	5056	4176
23	6319	6127	5832	5615	4969	4068
24	6258	6064	5765	5545	4890	3967
25	6202	6006	5704	5481	4816	3872
26	6150	5952	5647	5422	4748	3784
27	6102	5902	5595	5367	4685	3701
28	6057	5856	5546	5316	4626	3624
29	6016	5813	5500	5269	4570	3550
30	.5977	5773	5458	5224	4519	3481
40	.5695	5481	5149	4901	4138	2952
60	5414	5189	4838	4574	3746	2352
120	5133	4897	4525	4243	3339	1612
∞	4853	4604	4210	3908	2913	0000

INDEX